

The Effect of External Factors on Baseball

Kayoung Kim

duPont Manual High School

Abstract

The purpose of this project was to observe the relationship between five external factors in baseball with the number of runs scored for the Cincinnati Reds. Four of the external factors were related to weather conditions (temperature, relative humidity, air pressure, and air density), and the last external factor was the presence or absence of a full moon. With knowledge about these relationships, fans and teams in baseball could have a clearer understanding of the degree of the effect these selected external factors have on the score. Data was obtained through the NOAA. The four atmospheric factors were each separately compared with the home score of the Cincinnati Reds through correlation analysis. For the full moons, an uncorrelated, one-tailed t test assuming equal variances was performed. After all of the data was analyzed, there was no conclusive, statistically significant evidence to support the research hypothesis that air density and air pressure will have a negative correlation with the runs scored, humidity and temperature will have a positive correlation, and the presence of full moons will make the number of runs lower. However, when the external factors were examined more closely, trends in the data in the hypothesized direction, although weak, were discovered for three of the external factors (temperature, air pressure, and air density). To fully understand the relationship between baseball and its external factors, an expansion of the project using more data, more external factors, or a team batting statistic other than the number of runs scored is recommended.

Keywords: baseball, external factors, atmospheric, full moon

Introduction

Have you ever watched a baseball game or even just a part of it? It is likely that you have since it is the second-most popular sport in America, with American football being the only sport that beats its popularity (“Top 10 Most Popular Sports”, 2020). In baseball, when the pitcher throws the ball, the slightest change in the way he throws it or the slightest change in the way the batter hits it can have a huge impactful result on the game. These slight changes can be caused by simply the way the players throw or hit the ball, the players’ mental changes, or external factors such as the weather. The first cause is straightforward, as the game is naturally affected by the players’ performances. The second cause, the players’ mental changes, such as being nervous or not concentrated enough, can have a big impact also. Because of this, or due to a personal superstition, many players have their own unique ritual/routine that they do when it is their turn to either bat or pitch. Many baseball players over the years have had bizarre and complex rituals including steps such as adjusting their gloves, moving their feet in a certain way, or even licking the bat. The third, which is the external factors, can affect the game as well since baseball is a very sensitive sport. Changes in the weather or crowd noise are possible factors that could change the outcome. For example, a slight increase in humidity could be what turns a double or triple into a home run, helping the team win the game. The goal of this project is to investigate the relationship between external factors in baseball and the runs scored and examine if they show any type of correlation. Data from the Cincinnati Reds will be used to investigate this relationship.

This project will answer the following question: Does air density, its three factors considered separately (air pressure, humidity, and temperature), or full moons affect the number of runs scored in baseball? The first four possible external factors (air density, air pressure,

humidity, and temperature) are related to atmospheric conditions that would vary game by game. Atmospheric factors have been proven to affect how far a baseball travels once pitched and according to those studies, air density and the distance a baseball travels have an inverse relationship (Bahill et al., 2009; Branom, 2020). However, even though a ball may travel further during a hot, humid day, this does not necessarily mean that the number of runs scored in a game will actually increase, as there are a large number of other factors during a game other than just the distance of the flight of a baseball. No one has conclusive data to prove that these atmospheric factors affect the number of runs scored yet. The fifth possible external factor, the presence of a full moon, is chosen for investigation in addition to atmospheric factors as full moons may have an effect on baseball also. Full moons and their effects on humans is a controversial topic with many different opinions. Some people claim that full moons affect many aspects of human activity such as sleep cycles, human emotions, crime, and human aggression (Cajochen et al., 2013; Lieber, 1978a, 1978b), some disagree (Feltman & Kaplan, 2016; Foster & Roenneberg, 2018; Thompson & Adams, 1996), and some are undecided (Muñoz-Delgado et al., 2000; Parmar, 2014). If the lunar phases do affect humans, then the presence of a full moon affecting baseball games would have a high possibility as well.

The hypothesis for this experiment is: if the external factor is proven to show a positive correlation, then the runs scored in the baseball game will increase. All of the external factors (air density, air pressure, humidity, temperature, and full moons) are predicted to each show a correlation with, and therefore affect, the number of runs scored. Air density and air pressure are expected to have a negative correlation with the runs scored since higher density/pressure makes it harder for a ball to travel further. Humidity and temperature are expected to have a positive correlation since an increase in the levels of these will make it easier for the ball to travel further.

Finally, the presence of full moons is predicted to make the number of runs lower, since possible changes in human behavior, mood, or sleep cycles could lead to players feeling “off” and losing concentration or performing poorly.

The degree of impact will be observed by comparing the correlation coefficient from these atmospheric and astronomical factors to the correlation coefficient calculated from a team batting statistics (each one considered separately) and the number of runs scored. This will allow for the degree of impact of these external factors to be shown in a way that is clear and easily comprehensible by everyone. If the hypothesis is supported, the results can also be shared with baseball players, coaches, and spectators so that they can all be aware of these external factors and that they may affect the game they are about to play or watch. On the contrary, if this hypothesis is rejected, then baseball players can be assured that the game will solely depend on how well they perform, and superstitions such as the full moon affecting human behavior will be known as untrue for them. This will let baseball players focus on the game and not be distracted by possible external factors that could ultimately change the outcome. This will give them a more peaceful and concentrated mind so that they can do their very best during the games.

Methodology

The goal of this project was to observe the relationship between five selected external factors and the runs scored in baseball and to see if they strongly affect the game. This was done by gathering data from different websites and compiling them into an Excel spreadsheet. First, team batting statistics from Baseball-Reference for the Cincinnati Reds from seasons 2015 to 2019 were collected and inputted into the spreadsheet. The dates and scores were formatted into a month/day/year form and separated into the scores of the Cincinnati Reds, the scores of their

opponent, and the total score, respectively. The dates and the scores of the Cincinnati Reds were copied into another sheet in the same Excel spreadsheet. The scores of the Cincinnati Reds were chosen rather than the scores of the opposing team or the total score because the team batting statistics were from this team. Comparing the correlation between the external factors and the score of a team and the correlation between team batting statistics and the score of the same team makes the most sense.

Then, data about the temperature in Celsius, relative humidity, and air pressure in inches of mercury at 7:00 PM, since this is when most baseball games start, from the place the game was held were collected and inputted into the spreadsheet with the dates and scores. Data was collected using the NOAA Local Climatological Data Map Tool (see Appendix for link). Climate data in hourly CSV form was ordered and downloaded for dates 3/1/2015 to 12/31/2019 to ensure that the full season would be included. When finding the weather station to get the data from for each stadium, Google Maps was used to find the closest weather station to the stadium. If there was a tie, the geography (i.e. next to a river, next to the ocean) was taken into consideration and the most similar location was used. On average, each station was about 5 miles away from the stadium. The columns “HourlyDryBulbTemperature”, “HourlyRelativeHumidity”, and “HourlyStationPressure” were gathered from the downloaded data along with the date and hour. Then, the data was cleansed so that only the data for 7:00 PM (or closest to 7:00 PM after 7:00 PM) during dates when games were played remained. Finally, the data about moon phases from NOAA Tides & Currents (see Appendix for link) from 2015 to 2019 was collected and inputted into the spreadsheet. Only the dates when there was a full moon were collected since the other phases were of no importance to this project. The dates when there was a full moon and the dates that the Cincinnati Reds played in were cross-referenced so that

there was a “1” next to every game day with a full moon and a blank next to days without a full moon.

After all of the data was gathered, data cleansing was performed on each set and some outliers were taken out. Doubleheaders (days with two games) and suspended games were taken out completely since the meteorological data was collected for 7:00 PM and the times would change for one or both of the doubleheader games and the suspended games would not provide a complete representation of the game. Dates 4/30/2015, 6/15/2016, and 9/30/2017 were also taken out because there was either no recorded data from 7:00 to 8:00 PM unlike all other dates with data in that time period or because there was a source error for air pressure. Additionally, all games played in Toronto (5/29/2017, 5/30/2017, 5/31/2017) were taken out because NOAA did not have data for Toronto since it is outside of the United States.

Finally, one last external factor, air density, was added to the spreadsheet. The density of humid air was found by treating it as a mixture of ideal gases. The formula

$$p_{humid\ air} = \frac{p_d}{R_d T} + \frac{p_v}{R_v T} = \frac{p_d M_d + p_v M_v}{RT} \text{ where } p_d \text{ equals the partial pressure of dry air}$$

in Pascals, p_v is the pressure of water vapor in Pascals, and T is the temperature in Kelvin was used. When calculating the saturation vapor pressure for the vapor pressure of water, the Buck equation was used because this was the most accurate equation for the range of inputs that were present in this project. All of these calculations were performed by Excel after the formulas were inputted. After this, each row had a date followed by the air pressure at 7:00 PM, humidity at the same time, temperature also at the same time, air density, a one or a blank indicating the presence or absence of a full moon, and the score of the Cincinnati Reds.

Then, the final step of data analysis was performed. The four external factors (air pressure, relative humidity, temperature, and air density) were each graphed as a scatter plot with the number of runs scored by the Cincinnati Reds as the dependent variable (y-axis) and each of the external factors being separate independent variables (x-axis). The correlation coefficients of the scatter plots were found, then regression analysis was also performed to determine if there were any patterns unable to be explained by the correlation coefficient. For the full moon, a one-tailed uncorrelated t-test assuming equal variances with a significance level of 0.05 was performed with the scores of the days with a full moon being one group and the scores of the days without full moons being the other. In order to analyze the data, the Analysis Toolpak add-in was first added to Excel so that the tests could be performed. Afterward, before the t test, a one-tailed F test was performed first to see the variance of these two groups. The null hypothesis was that the two groups will have the same distribution and therefore not display a difference, and the alternate hypothesis was that a difference is displayed. A t test was performed to see if it could reject the null hypothesis.

After the analysis of all the external factors, correlation analysis was done for a team batting statistic, the Total Average. This statistic was found by calculating using different team batting statistics provided by Baseball Reference (see Appendix for link) in the equation $TA = \frac{TB+BB}{AB-H}$ where TA is the Total Average, TB is the number of Total Bases, BB is Base on Balls, AB is At Bats, and H is Hits. The correlation coefficients were compared to see to what extent the external factors affect the game, keeping in mind that it may not be to the extent of the team batting statistics that are a part of the game.

Data and Results

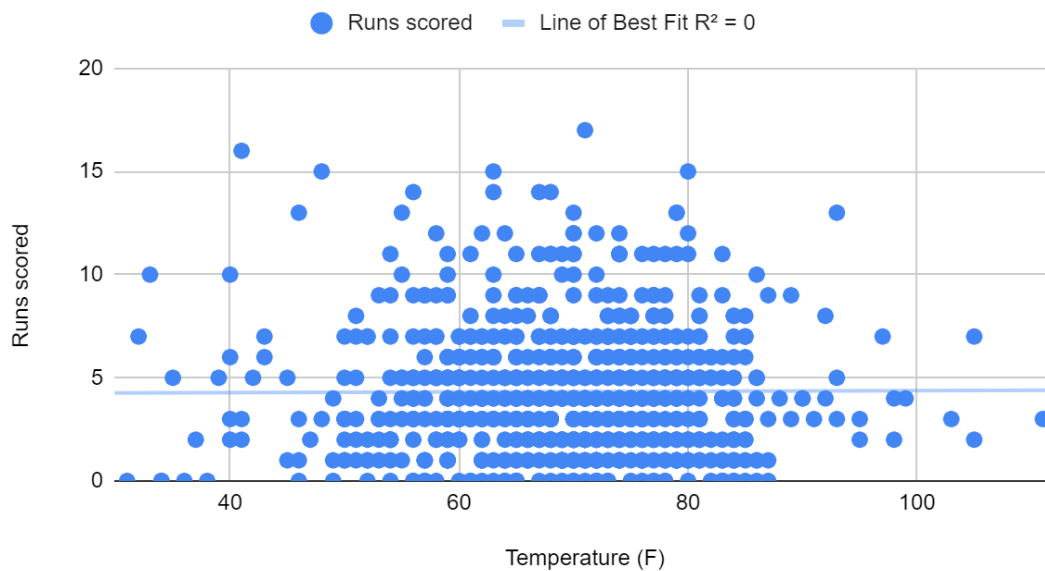
Data gathered from the multiple sources (see Appendix for the full list) were organized into an Excel spreadsheet. The raw data collected was the dry bulb temperature (F), relative humidity (%), air pressure (Hg), and whether or not there was a full moon. The table below (Table 1) is one row of the full raw data table (see Appendix for full raw data) for the external factors. The last external factor, air density, was calculated using these raw data values.

date	place	Temperature (F)	Humidity (%)	Air Pressure (Hg)	moon
6-Apr	CIN	55	100	29.52	

Table 1. Sample raw data table for external factors.

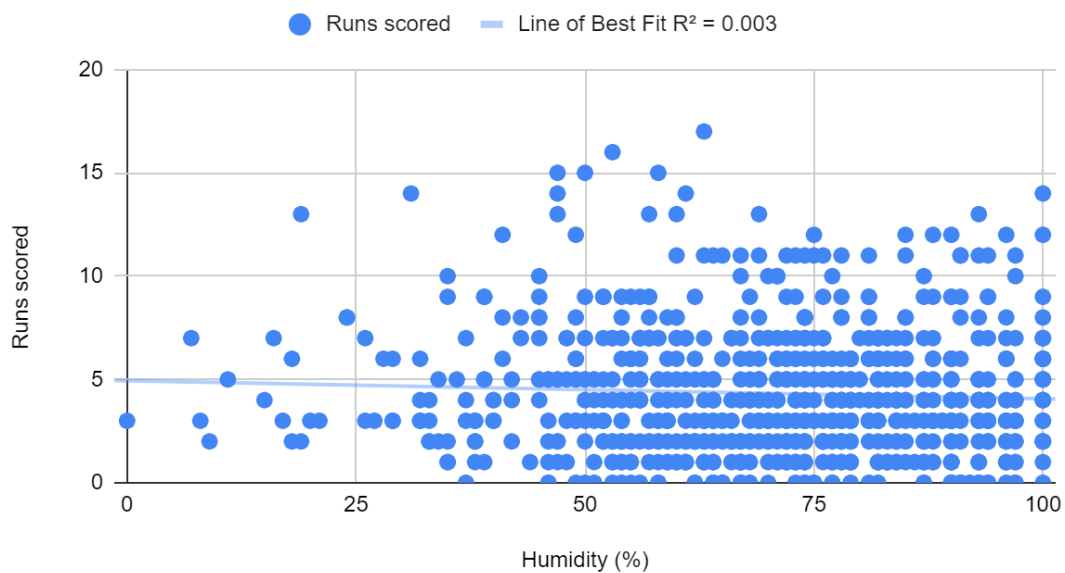
Each of the five external factors was analyzed in comparison to the number of runs scored by the Cincinnati Reds. The four external factors related to the weather: temperature, humidity, air pressure, and air density, were graphed each with the runs scored on a scatter plot (see Graphs 1, 2, 3, and 4). Correlation analysis was performed for each graph and the correlation coefficient was calculated. The correlation coefficients for each of the graphs were calculated as the following: 0.005824322 for temperature, -0.053159915 for relative humidity, -0.126041456 for air pressure, and -0.095936452 for air density.

Runs scored vs. Temperature (F)



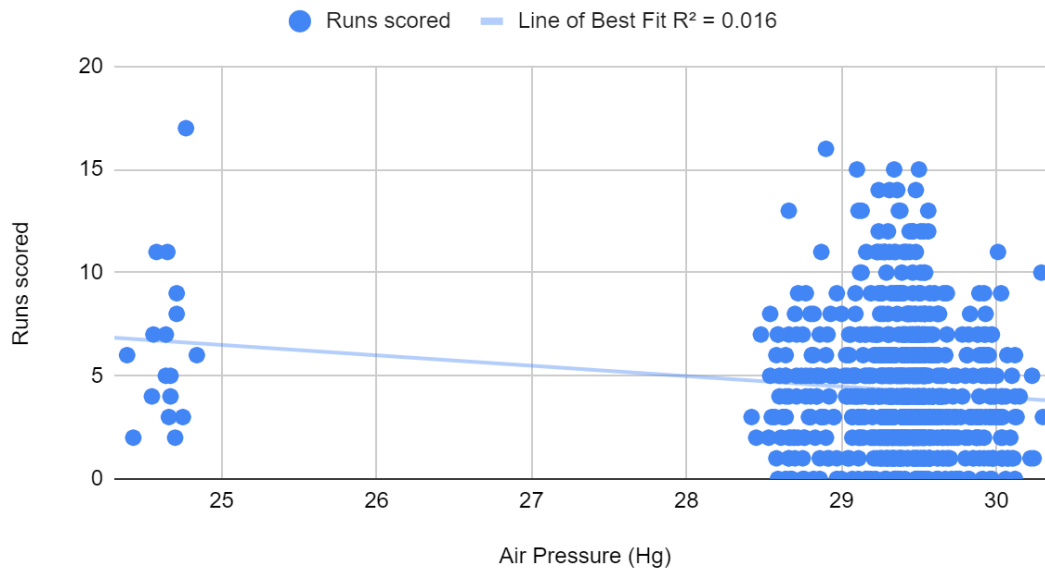
Graph 1. Displays the relationship between temperature (IV) and the number of runs scored (DV) during seasons 2015-2019 for the Cincinnati Reds.

Runs scored vs. Humidity (%)

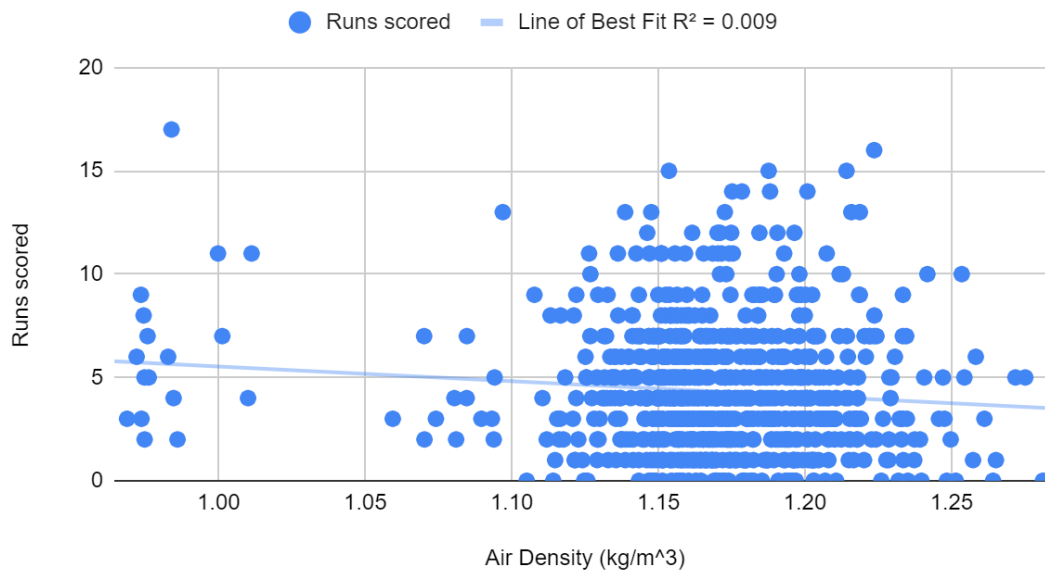


Graph 2. Displays the relationship between humidity (IV) and the number of runs scored (DV) during seasons 2015-2019 for the Cincinnati Reds.

Runs scored vs. Air Pressure (Hg)



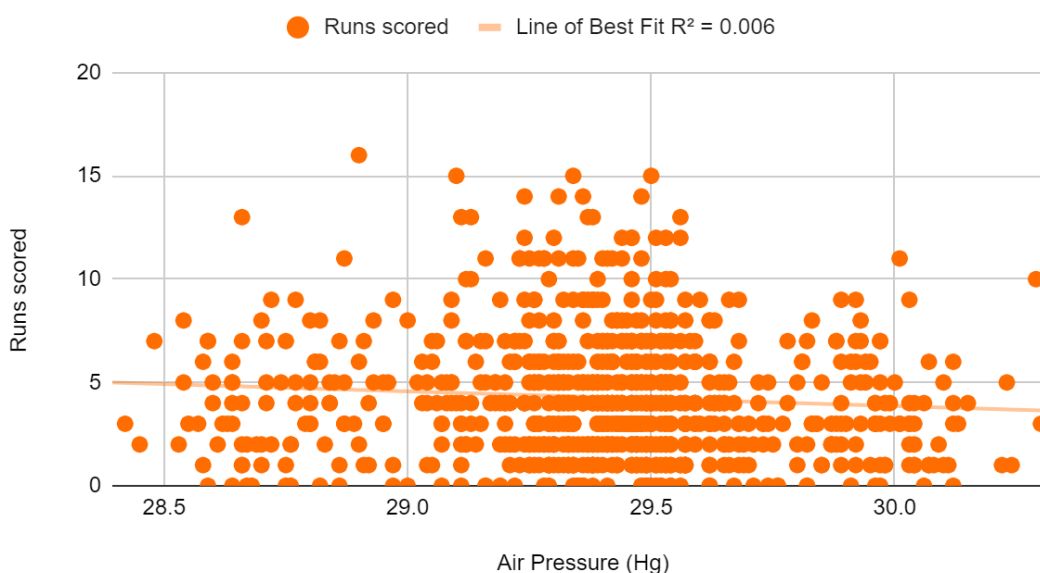
Graph 3. Displays the relationship between air pressure (IV) and the number of runs scored (DV) during seasons 2015-2019 for the Cincinnati Reds.

Runs scored vs. Air Density (kg/m^3)

Graph 4. Displays the relationship between air density (IV) and the number of runs scored (DV) during seasons 2015-2019 for the Cincinnati Reds.

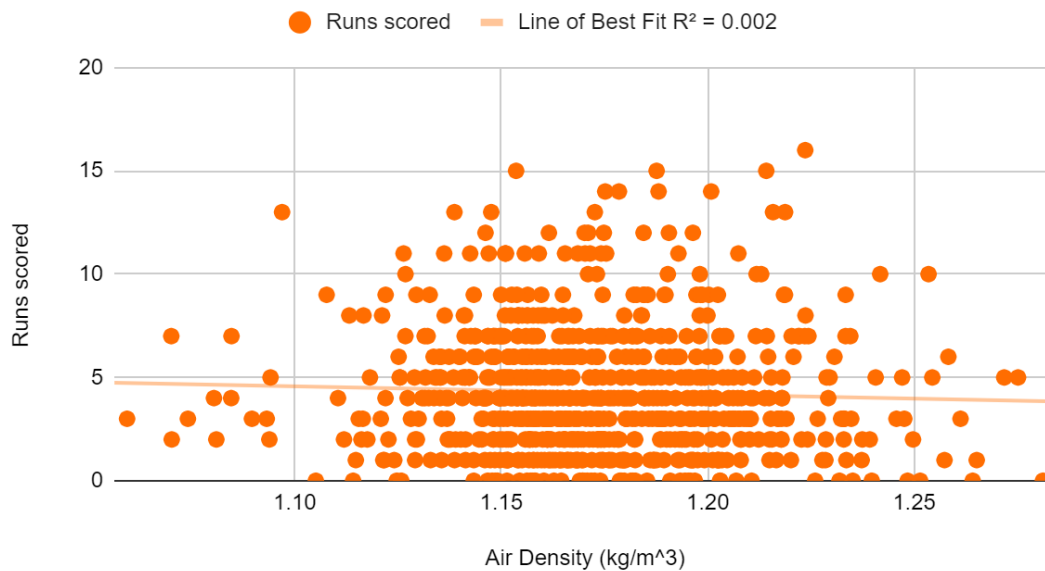
In graphs 3 and 4, it is very clear that the few points plotted on the left side of the graph are very different from the rest of the points on the graphs. These few and the rest of the points have a large horizontal gap between them. When the data plotted was examined, it was found that one of the locations that the Cincinnati Reds played at had abnormally low air pressure. This was the stadium of the Colorado Rockies, and Colorado's very high elevation had affected the air pressure significantly. Additionally, since the air density is affected by the air pressure in addition to temperature and humidity, the aforementioned gap in the graph can be seen, although not as extreme, as air density has the other factors mentioned above. Because these few points were impacting and distorting the overall line of best fit and correlation coefficient, two additional scatter plots were created to assess the impact of air pressure and air density on the score for the rest of the less extreme locations (see Graphs 5 and 6). The correlation coefficients for these two graphs were -0.079529926 and -0.040913404 for air pressure and air density, respectively.

Runs scored vs. Air Pressure (Hg)



Graph 5. Displays the relationship between air pressure (IV) and the number of runs scored (DV) during seasons 2015-2019 for the Cincinnati Reds WITHOUT the values from Colorado.

Runs scored vs. Air Density (kg/m^3)



Graph 6. Displays the relationship between air density (IV) and the number of runs scored (DV) during seasons 2015-2019 for the Cincinnati Reds WITHOUT the values from Colorado.

The magnitudes of the correlation coefficients for air pressure and air density are comparatively greater than the rest and are in the hypothesized negative direction. Temperature is in the hypothesized positive direction but the magnitude of the correlation coefficient is very low. Relative humidity is in the opposite direction of the hypothesis but the magnitude is also very low.

The last external factor, the presence of full moons, was analyzed with a t test with the first group being the home scores of games without a full moon present and the second being the scores of games with full moons. Before the t test, an F test was conducted first to see if they had

equal variances (see Table 1). Because the F value (1.053859935) is less than the critical value (1.700170089), the null hypothesis that the variances were equal was not rejected.

	<i>No full moon</i>	<i>Full moon</i>
Mean	4.339498018	4
Variance	9.565805568	9.076923
Observations	757	27
df	756	26
F	1.053859935	
P(F<=f) one-tail	0.461739913	
F Critical one-tail	1.700170089	

Table 2. One-tailed F test to test variances and measures of central tendency.

The measures of central tendency and variability were observed next. The mean of the first group without full moons was 4.339498018 and the mean of the second group with full moons was 4. The medians of both groups were 4. The variance for the first group was 9.565805568 and for the second group, it was 9.076923077. The two groups have similar means, medians, and variances, suggesting that the two samples are likely to have come from the same population. A one-tailed uncorrelated t test assuming equal variances was conducted with the level of significance being 0.05. The t value given by the t test was 0.560941478, and the critical value was 1.646804505. Since the t value was lower than the critical value, the null hypothesis that the samples come from the same population was not rejected (see Table 3).

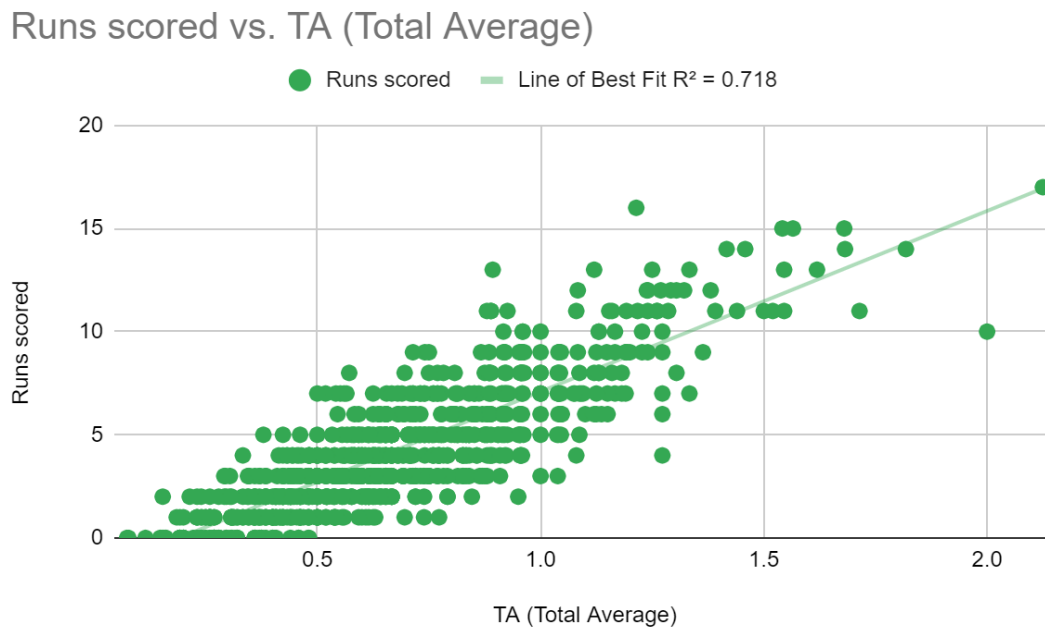
	<i>No full moon</i>	<i>Full moon</i>
Mean	4.339498018	4
Variance	9.565805568	9.076923
Observations	757	27
Pooled Variance	9.549551163	
Hypothesized Mean Difference	0	
df	782	
t Stat	0.560941478	
P(T<=t) one-tail	0.287499061	
t Critical one-tail	1.646804505	

Table 3. One-tailed uncorrelated t test assuming equal variances to test if the samples are from the same population and measures of central tendency.

After analyzing the external factors, team batting statistics were analyzed to see the level of influence they have on the number of runs scored. The Total Average (TA) was calculated using the raw data gathered shown by a sample row below in Table 4 (see Appendix for full raw data). The TA was graphed with the runs scored (see Graph 7). After correlation analysis was performed, the correlation coefficient was found to be 0.847555032. This is much greater than the correlation coefficients of the graphs for the three atmospheric external factors and the astrological external factor.

Date	AB	H	2B	3B	HR	BB
6-Apr	29	5	0	0	2	3

Table 4. Sample of the team batting statistics raw data for calculating the Total Average (TA).



Graph 7. Displays the relationship between the TA (IV) and the number of runs scored (DV) during seasons 2015-2019 for the Cincinnati Reds.

After all of the data for each external factor was analyzed and compared with the team batting statistics, strong, conclusive evidence to support the research hypothesis that air density and air pressure are expected to have a negative correlation with the runs scored, humidity and temperature are expected to have a positive correlation, and the presence of full moons is expected to make the number of runs lower was not found.

Conclusion

The purpose of this project was to observe the relationship between five external factors in baseball with the number of runs scored for the Cincinnati Reds for five seasons from 2015 to 2019. Four of the external factors were related to weather conditions (temperature, relative humidity, air pressure, and air density), and the last external factor was the presence or absence of a full moon. The four atmospheric factors were each separately compared with the home score

of the Cincinnati Reds through correlation analysis. For the full moons, a t test was performed to compare the home scores when there was a full moon present and those when a full moon was not present.

After all of the data was collected and analyzed, it was concluded that the research hypothesis, which states that air density and air pressure are expected to have a negative correlation with the runs scored, humidity and temperature are expected to have a positive correlation, and the presence of full moons is expected to make the number of runs lower, was not supported by the project. The t value for the full moons t test was 0.560941478, and the critical value was 1.646804505, meaning that the null hypothesis was not rejected, and the correlation coefficients were the following: 0.005824322 for temperature, -0.053159915 for relative humidity, -0.126041456 for air pressure, and -0.095936452 for air density. When compared to the correlation coefficient for a team batting statistic, the Total Average, which was 0.847555032, it can be concluded that none of the external factors had a significant impact on the runs scored.

When each of the atmospheric external factors is observed separately, interesting trends in the data can be seen. It can be observed that air pressure and air density show the predicted effect on the score with comparatively more clarity than the others. Because objects travel slower at higher levels of air pressure and air density, the negative correlation coefficients make sense and support the science behind it. The temperature had a positive correlation coefficient, as expected, but the magnitude was extremely low. When air warms, it expands, which lowers the density of the air. This produces longer flight distances, assuming that other factors stay the same. However, because the correlation coefficient is only slightly positive, it is not enough to draw a solid conclusion about the relationship. The same trend can be observed for the relative

humidity. The correlation coefficient for this factor was slightly negative, which is the opposite of what the research hypothesis predicted since air is less dense when there is more water vapor present. However, once again, since the magnitude of the correlation coefficient is very low, the relationship between the humidity and the runs is not clear.

One may wonder why the results of correlation analysis for each of the weather-related external variables and the runs scored did not completely produce the desired results that would be logical according to the scientific principles related to it. One possible source of error is that the stations from which the weather data was collected may not be exactly where the game took place. Since the NOAA Local Climatological Data tool was used, the stations available with weather data were not exactly at the stadium where the game was played. The closest station, which was always used for each location, was on average about 5 miles away from the stadium. Although this distance difference is not large, the weather certainly may have been a bit different at the actual stadium than at the station at which the data was collected. Another possible source of error is that the game may not have been played during the hour that the experiment used to collect data for the atmospheric factors. 7 P.M. (in the time zone of the game location) was used for all games to collect the weather data since 7 P.M. was a reasonable estimate that would be near the start of the game. However, there may have been games that were played earlier or later due to special circumstances specific to that game. There may also have been weather changes after the game started (after 7 P.M.) that had an impact on the game overall. For example, at the beginning of the game around 7 P.M., it may have been partly sunny, but about halfway through, it is possible that the weather changed and started raining instead.

In order to truly further understand the relationship between each of the external atmospheric factors and the runs scored, further research is necessary. One possible extension is

to simply look at more data, either with more teams, longer time periods, or both. This is because the bigger the sample size is, the better it represents the actual population. Also, when only the Cincinnati Reds were observed, their home stadium was used very often compared to the other stadiums during away games which may have altered some outputs, and each team has their own unique assets and challenges which may have interfered as well. Another way to further explore the relationship between external factors and baseball is to compare the external factors with another statistic in baseball other than the score. Some good examples include the Slugging Percentage (SLG), how far the ball travels, and the number of home runs. These statistics in baseball may represent the effect of external factors better than the score since the score is the final value that has numerous factors including the external factors, team batting statistics, how well the other team plays during that specific game, and more. Focusing on just a specific statistic may reflect the complex relationship better. Another good extension of this project would be to observe additional external factors, such as social media. The full moon was related to the baseball players' minds and emotions and the possible psychological effects of full moons, and observing social media influences would further dive into external factors that affect the players themselves. Usually, before the game or before the start of the season, there are many people on the Internet discussing the possible results and events that would occur. There may also be ongoing rumors or scandals about the team and their reputation. These discussions on social media may affect the players' concentration and mind, leading to a bad result for the team. Investigating more external factors such as this one will help players, coaches, and fans of baseball better understand the complex sport that they all enjoy and love.

As of now, however, by looking at the results for this project, it would be very reasonable to inform fans and teams that external factors, although they do have an effect, and some more than others, can be disregarded when watching a game. In the end, the game itself, and both the subtle and exciting moves that occur during the game, are what really affect the runs scored in baseball.

Appendix

NOAA Local Climatological Data Map Tool: <https://www.ncdc.noaa.gov/cdo-web/datatools/lcd>

NOAA Tides & Currents: <https://tidesandcurrents.noaa.gov/astronomical.html>

Baseball Reference:

<https://www.baseball-reference.com/teams/tgl.cgi?team=CIN&t=b&year=2015>

Full Raw Data:

<https://docs.google.com/spreadsheets/d/15LjZcnfbgQm8UggXbKJ6W19d7nOAQqSxseMCAvVFGPs/edit?usp=sharing>

References

Bahill, T., Baldwin, D., & Ramberg, J. (2009, February 5). Effects of altitude and atmospheric conditions on the flight of a baseball. Academia.edu - Share research.

https://www.academia.edu/8605937/Effects_of_Altitude_and_Atmospheric_Conditions_on_the_Flight_of_a_Baseball

Branom, M. (2020, October 4). Fair skies, foul balls: How weather influences baseball. The Washington Post.

<https://www.washingtonpost.com/weather/2020/10/04/weather-influence-baseball/>

Cajochen, C., Altanay-Ekici, S., Münch, M., Frey, S., Knoblauch, V., & Wirz-Justice, A. (2013, July 25). Evidence that the Lunar Cycle Influences Human Sleep. Current Biology.

[https://www.cell.com/current-biology/fulltext/S0960-9822\(13\)00754-9](https://www.cell.com/current-biology/fulltext/S0960-9822(13)00754-9)

Feltman, R., & Kaplan, S. (2016, October 24). Dear Science: Does a full moon really change human behavior? The Washington Post.

<https://www.washingtonpost.com/news/speaking-of-science/wp/2016/10/24/dear-science-does-a-full-moon-really-change-human-behavior/>

Foster, R. G., & Roenneberg, T. (2008, September 9). Human responses to the geophysical daily, annual and lunar cycles. ScienceDirect.com | Science, health and medical journals, full text articles and books.

<https://www.sciencedirect.com/science/article/pii/S0960982208008658>

Lieber, A. L. (1978). The lunar effect: Biological tides and human emotions. Doubleday.

Lieber, A. L. (1978). Human aggression and the lunar synodic cycle. Journal of Clinical Psychiatry. <https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=49286>

Muñoz-Delgado, J., Santillán-Doherty, A. M., Mondragón-Ceballos, R., & Erkert, H. G. (2000, October 3). Moon cycle effects on humans: myth or reality? Medigraphic - Literatura Biomédica. <https://www.medigraphic.com/pdfs/salmen/sam-2000/sam006e.pdf>

Parmar, V. (2014, January 12). Effects of full-moon definition on psychiatric emergency department presentations. Publishing Open Access research journals & papers | Hindawi. <https://www.hindawi.com/journals/isrn/2014/398791/>

Thompson, D. A., & Adams, S. L. (1996, March). The full moon and ED patient volumes: Unearthing a myth. ScienceDirect.com | Science, health and medical journals, full text articles and books. <https://www.sciencedirect.com/science/article/abs/pii/S0735675796901242>

Top 10 Most Popular Sports in America 2020 (TV Ratings). (2020, October 2). <https://sportsshow.net/most-popular-sports-in-america/>.