Altitude and Its Effect on MLB Park Factor

Kevin Nguyen
University of Colorado Boulder
Boulder, CO
Keng3875@colorado.edu

Samuel Busser
University of Colorado Boulder
Boulder, CO
sabu0818@colorado.edu

Blaise Page
University of Colorado Boulder
Boulder, CO
blpa1200@colorado.edu

Phillip Arsenian
University of Colorado Boulder
Boulder, CO
phar0872@colorado.edu

1.1 ABSTRACT

The main question we are seeking to answer is whether or not altitude really has a significant effect on the game of baseball. We will answer this question by looking at offensive statistics such as batting average, runs per 9 innings, the ratio of home runs to total runs, and several more offensive stats.

After our investigation, we concluded that altitude does not have a direct impact on the game of baseball. Of the offensive statistics that we looked into, the only ones that had a significant correlation with altitude were batting average and singles. However, this correlation most likely does not show causation. For one, these stats were the only ones that had a significant, positive correlation with altitude.

1.2 INTRODUCTION

Major League baseball is a very competitive sport where every year 30 teams compete to prove that their team is the best at the sport in the country. One of the most important factors in sports competitions is to ensure that the game is played fairly with no one team having an advantage over the other. Altitude is one of the biggest components considered when judging the fairness of sport competitions because it certainly has an effect on almost all sports. However, with most sports, altitude is mostly expected to affect only the endurance of players who are not accustomed to playing at such high elevations where oxygen is not so readily available. Baseball is the only sport where altitude can actually have an effect on the travel path of balls hit which can make a big difference in the performance of a team. This can certainly provide an unfair advantage to teams that play at higher elevations.

Coors Field in Denver is located 5,183 feet above sea level (A mile high at the top of the stadium). This is by far the stadium with the highest altitude in the MLB. The stadium with the second highest altitude is Chase Field which is located 1,082 feet above sea level. This altitude difference is quite alarming especially if altitude does have an effect on the performance of a team. In order to combat the unfair advantage of altitude, the Colorado Rockies have been storing their game used baseballs in a humidor since 2002. This was meant to prevent the ball from drying out which would consequently make the ball bouncier and allow it to travel farther after being hit. However, it has never been proven that keeping the ball humid has helped prevent the Colorado Rockies from having an unfair advantage on their home field. Through the use of dating mining techniques, we wish to expose whether or not players perform better at higher altitudes.

1.3 PROBLEM STATEMENT/MOTIVATION

Since its creation in 1993, Coors field has notoriously been one of the most hitter friendly parks in the major leagues due to the high altitude. The Rockies have tried to combat this by storing baseballs in a humidifier. However, there are several parks across the league that are also considered hitter friendly that are not at a high elevation.

Our goal is to look at teams and players stats throughout the stadiums played in the MLB (pitching, hitting, and fielding) and try to determine how big of an impact altitude has on players stats and the game of baseball in general.

Figure 1: Humidors in Coors Field to keep the baseballs from getting dry

2 RELATED WORK

This idea of the park affect at Coors Field and many other stadiums across the league has been debated for a long time, thus there is lots of work on this topic already. However, most of this work deals with teams at a general level, and does not consider the effects of altitude specifically.

$$PF = \frac{\left(\frac{Run\ Scored(Home)+Runs\ Allowed(Home)}{Number\ of\ Games\ at\ Home}\right)}{\left(\frac{Run\ Scored(Road)+Runs\ Allowed(Road)}{Number\ of\ Games\ on\ theRoad}\right)}$$

Figure 2: Current Park Factor Formula

The MLB has instituted a statistic called park factor. However, this statistic does not focus on the altitude of the stadium or anything local to the stadium in general. Instead, it is calculated by comparing a team's runs scored in games at home versus a team's runs scored in games on the road. We will differentiate from this, as we will focus on team's offensive, defensive, and pitching performances at different altitudes.

Since the turn of the century, statistics and data analytics has become an increasingly important aspect of the game of baseball (famously coined "moneyball"). This kind of thinking has revolutionized the game, but again the work done in moneyball is not the work we will be doing. The moneyball revolution emphasized on base percentage and prioritizing players who get on base in any way possible. While we may use on base percentage as one of our statistics to evaluate the difference altitude makes in the game, we will not solely be looking at this one stat.

One study published in 2008 used regression analysis to model the effects of four factors (altitude, temperature, humidity, and barometric pressure) on the flight of a baseball or softball under different atmospheric conditions [Bahill, A. Terry & Baldwin, David & S Ramberg, John, 2009]. The study was able to show that on a typical July afternoon, altitude clearly has the biggest impact on ball flight accounting for 80% of the variability between ball flight in different stadiums. The study directed most of its attention towards the force acting opposite of the relative motion of a moving object which is known as the object's drag force. The study determined that the drag force of a ball travelling through the air is weakest in Denver and strongest in San Francisco due to the altitude differences of both cities. This would suggest that if two balls were to be hit with an identical force and direction in both Denver and San Francisco, the ball in Denver would travel further under similar weather conditions. This conclusive information is already a good indication of altitude certainly providing an unfair advantage for teams playing at higher elevations. However, altitude is not the only factor affecting the ball flight of a batting hit for a given stadium. In fact the study also mentions that Coors field, which is the stadium with the highest altitude, has a high probability of experiencing high winds opposite to the direction of typical ball flight in a game. This factor decreases the number of home runs in the stadium and nearly compensates for the stadium's high altitude. Therefore, it is extremely hard to determine if altitude provides teams with an unfair advantage in terms of ball flight in their home stadiums because of the all the other factors that may also affect the travel flight of a baseball in each stadium. For this reason, we wish to focus most of our attention on the statistics recorded for each MLB game between the years 2005 and 2015. Of course, we will also have to account for the different variables that can affect the travel flight of a ball in each stadium. However, our conclusions can certainly be used to either support or refute previous work done on the altitude effects of different stadiums in the MLB.

Park Factor Stat:
http://www.espn.com/mlb/stats/parkfactor

Moneyball: The Art of Winning an Unfair Game by Michael Lewis

Bahill, A. Terry & Baldwin, David & S Ramberg, John. (2009). Effects of altitude and atmospheric conditions on the flight of a baseball. International Journal of Sports Science and Engineering.

3 PROPOSED WORK

Data Collection: Each member in our group has downloaded the dataset from the retrosheet database on our respective machines. Link to database: http://www.retrosheet.org/

Preprocessing: Our project requires us to omit all attributes and events that are not useful for determining how altitude affects the offense in a ballpark. These attributes include (but are not limited to) For example, a steal is an example of a specific event in the retrosheet database. This type of event will not be helpful for our project, so we are safe to remove these events. We can also do this for defensive interference, caught stealing, pickoffs, wild pitches, passed balls, balks, and several more. In addition to removing specific events, we can also remove certain attributes from each event that we do keep. Each event has a large number of attributes, many of which are much too specific to be useful for us. Some examples of these attributes are the handedness of the pitcher and hitter, what players were playing each position in the field, which players were on base, what player was on deck, and many more. We also have to account for NULL entries in the dataset. These are very few and far between, as RetroSheet has done a good job of providing a complete database. Because there are so few NULL entries, we can simply remove these entries and not have to worry about losing significant data. Doing all of this will help our results look much cleaner and easier to handle. More importantly though, it will greatly improve the speed of any calculations we will be doing with the data.

Process for derived information: As we have stated, Retrosheet contains every event from every game. However, it does not contain player or team statistics. This is where the bulk of our project lays. We will have to go through every event and compute the statistics ourselves. For example, if we are looking for all the home runs that occurred at Coors Field, we would have to go through every event that had the event code of a home run, and had a home team ID of Colorado. We will have to do this for every statistic that we wish to compute.

Design: We plan on collecting offensive, defensive, and pitching statistics for each ballpark, along with the elevation of each park. One statistic we will use will be home runs. This is a big one, as higher elevations are supposed to make the ball fly further, thus leading to home runs. We will also calculate the total number of runs scored in each park on average, as this is also a good indicator of the total offense in a game. We will also look at pitcher's combined ERA at each park. This will give us an idea of how many runs are allowed by pitchers. At high altitude, balls will not break as early as at low altitude, which can lead to a lot of easy to hit pitches. Fielding percentage is a defensive statistic we will look at. High altitude helps the ball fly further but also faster. Due to this increase in speed, it can be harder to field the ball, leading to more errors and lower fielding percentage. These are just a few of the statistics that we will calculate.

Evaluation: Once we have all the statistics we need, we can dig into our problem statement. To first prove that altitude has any affect at all on baseball games, we can perform a hypothesis test, with our null hypothesis being: Altitude has no effect on the game, and our alternate hypothesis being: Altitude has an effect on the game. If we are able to reject the null hypothesis, then we can confirm altitude does effect the game. We can then use correlation analysis to determine which stats are most affected by altitude.

There could be several uses for this information if we do find a correlation between altitude and some of the stats. For example, if someone was looking to create a new team, they could consult our findings and choose a city with an according altitude. Teams could also use this information when drafting players. If, for example, we find that home runs and altitude are very strongly correlated, then a team like the Rockies should look for good home run hitters, as they play half their games at a very high altitude.

4 DATA SET

Our dataset is from the Retrosheet database. This database contains every single play that has occurred in every MLB game since 1989. In addition to what happened on each play, this database also contains how many outs there are, how many runners are on base, how many pitches were thrown in the at bat, what specific kind of pitch was used, where the ball was hit, who was at each fielding position, and much more.

Our dataset for the MLB events spans between 2005 and 2015 seasons. Our dataset contains approximately 2,111,526 usable objects, which include both players and teams. The dataset also contains 41 attributes for which the corresponding team/player statistics are generated.

The attributes we used were the GAME ID, EVENT_CD, HOME_TEAM_ID, and EVENT_RUN_COUNT. GAME_ID was just the ID of each game, HOME_TEAM_ID allowed us to see what stadium the game was played at, EVENT_RUN_COUNT gives how many runs were scored during each play, and EVENT_CD tells us what happened during the event such as a hit, or strikeout.

5 EVALUATION METHODS

Our evaluation method will begin with comparing the altitude of each ballpark with the statistics we get from each ballpark We will see if there is a correlation between the two. We will also attempt to use a null hypothesis to see if all the stadiums are similar in statistics, and we will test this using confidence intervals to see if we can reject the null hypothesis or fail to reject it.

6 TOOLS

The database that we will be using to derive our conclusions contains every single play in the MLB since 1989 including a variety of attributes that won’t be helpful in exploring the park factor of a team. For this reason, one of the first tools we will be using is MySQL Workbench which will allow us to edit our database to consider only data that would be helpful in researching the effects of altitude in MLB statistics. This first step will make subsequent steps more manageable by eliminating a lot of the noise in our dataset and allowing us to more easily focus on the data that is more important for our project.

Our primary tool for our project will be a Jupyter Notebook which will allow us to manipulate our data and perform calculations through the use of python code and libraries. In order to easily manipulate our data, we will be using the PANDAS library. This library will allow us to smoothly load our dataset onto our Jupyter Notebook and will handle parsing our data into rows and columns. For most of our calculations, we will be using the NumPy library which offers additional functionality for large multidimensional arrays. Using NumPy supported functions will make our statistical analysis of our data much easier by allowing us to easily calculate important properties of our data such as the mean, standard deviation, and variance of specific attributes of our data. We will also use NumPy to run correlation tests and create confidence intervals to support our conclusions. We will be using matplotlib to visually represent our data through the use of different visualization tools such as boxplots, scatterplots, and histograms. These visualization tools will help us better support and explain our findings. \

7 MAIN TECHNIQUES APPLIED

Data Preprocessing: Our data was over 2 million data points alone, so we had to use many techniques we learned with preprocessing to make our data more easily accessible and usable. These are the techniques we used:

Data Cleaning: Since our data has spanned from 2005 - 2015, we had an instance where we had a team that moved cities, so we had to combine the statistics for Florida and Miami since they were the same team but In different stadiums. We had to make sure the altitude at both stadiums were equal because if not they would be played at different altitudes.

Data Reduction: Since our data was so large, we had to go through our data and remove columns that we did not need, such as who was on base during the hit, or what the pitch count was. In the end we only kept the columns that were included in the statistics we used, the home team the game was played at, and the game ID. We also had to group by stadiums for some statistics that had to do with totals at a stadium such as total hits, and total runs.

Data Integration: Since Retrosheet did not have an altitude column, we had to add that to the data. We created a table that included the home team and the altitude that the game is played at. We then merged Retrosheet and our table on the home team ID and that allowed us to add altitude to each event.

Correlation: Once we got our data for each statistic, we then looked for a correlation between the data and the altitude the games were played at. This was to see if altitude can affect baseball statistics or not.

8 MILESTONES

Our first step in this project will be to get the data on each one of our devices. We have completed this milestone as of March 1st.

Our next step will to be clean the data, and remove all unnecessary columns and data. We will want this done by March 12th. Once we have that done, we will need to combine the events of each game and have them organized. This should be done by March 19th.

Lastly, we will have all the data we need, so now we need to run our correlation tests, and confidence intervals to get all the comparisons we need. This will need to be done by April 13th, and once that is done we will need to conclude on our data and evaluate what it means. This will be done by April 30th.

8.1 MILESTONES COMPLETED

Our first milestone of getting the data on all of our machines has been completed. We first had to download the .csv file. Once this was done, we then integrated this data into a dataframe in Pandas and a table in MySQL.

We have also finished cleaning our data, at least for now. Depending on where the project goes, we may end up removing more attributes or rows, or adding more back in. For now, we have removed several columns from the database that will not be needed for our project. Some examples are time_since_1900 (measures the time since 1900 to that game), game_id, bat_hand_cd (simply whether the batter was left or right handed), base runner IDs, pitch sequence transactions, ball and strike count, and more.

Some of those are pretty self-explanatory as to why we won't need them. Some of them, however, could prove useful in other projects, but not for ours. For example, we don't need whether a hitter was left or right handed, as we are concerned with offensive statistics as a whole.

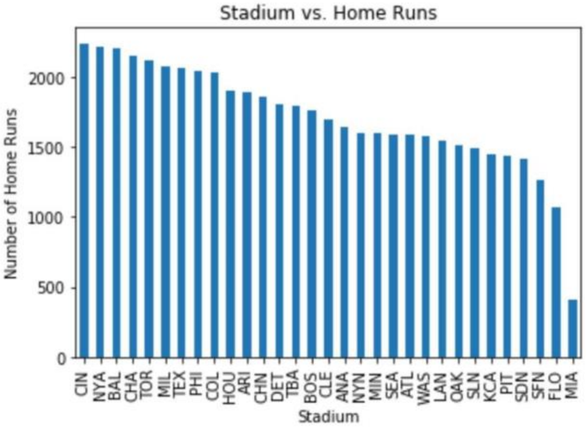
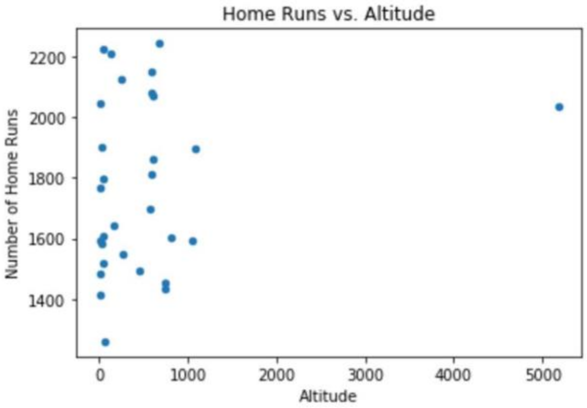


Figure 3: A bar chart showing number of home runs hit in each stadium

We also added the corresponding altitude of the location of each event in the database. By doing this, we will now be able to create graphs and charts comparing offensive stats to altitude. Because Retrosheet does not provide us with specific stats for each team, we have begun calculating offensive stats for each team as well. We are able to do this through the event code column in Retrosheet. The event code column has a number corresponding to what exactly happened in the at bat. So, for example, we have calculated the number of home runs that occurred at each stadium, as well as the number of total runs.

8.2 MILESTONES TO DO

One big milestone that we have left to do is to officially decide which statistics we are going to use in showing whether there is a correlation between offensive production and altitude. So far, we know we want to use total runs scored and total home runs at each stadium. But this is just a top level look. There is much more we can look at.



For example, it has been shown that ground balls at stadiums with higher altitude will have a faster velocity. We could use this piece of information to compare the total number of errors at stadiums of high and low elevation. Additionally, it has been proved that pitchers cannot get a ball to break as much at higher elevations. Because of this, if they try to throw some form of curve ball, it might not curve as much as they are used to, leaving the ball in a better spot for the hitter to hit. This could lead to pitchers having worse pitching statistics at stadiums of higher elevations as well. So ERA and runs per 9 innings are two more stats that we will look at in the future.

Another future milestone we have is to integrate park factor into our dataset. While the park factor stat that the MLB has implemented does not use the same information we have, it would still be interesting to compare our results to results that have been found using a different method.

Finally, our last significant milestone remaining is to perform all of the data science correlation calculations on our data to see if there really is a correlation between offensive production and altitude. This will include hypothesis tests, computing correlation coefficients, and r-squared values.

9.1 RESULTS SO FAR

With the data set we have cleaned so far, we have been able to create different visual representations to compare the altitude of each stadium with the statistics of that stadium in between the years 2005 and 2015. The first diagram we created was a bar chart that mapped the frequencies of home runs to each teams' home stadium. This bar chart allows us to visualize the variations in the number of home runs scored at each stadium. While this diagram only shows home runs, it gives us some insight into our problem already. Before we saw the results, one would expect Coors field to have the most home runs since it is at the highest elevation. As can be seen in the graphs below, Colorado is in the top ten, but it does not have the most home runs hit. The team that had the most home runs hit in their stadium was the Cincinnati Reds, with the New York Yankees and Baltimore Orioles close behind. Interestingly, the elevation of these three cities are 482, 26, and 36 feet respectively. These are all miniscule compared to the mile-high elevation at Coors Field, and yet they all have a higher home run rate than Coors Field does. Again, this is only one stat that we have looked at so far, so it is not conclusive. But home runs are a pretty significant stat in the altitude park factor argument, so the fact that the highest altitude does not have the highest park rate is very interesting.

Our next step was to create a diagram that would help us better isolate the effects of elevation on the number of home runs hit at different altitudes. Therefore, we created another bar chart that compared the number of home runs hit at specific elevations. From this graph we were able to see that most the home runs were definitely hit at regions of higher elevations. However, the impact of altitude on the number of home runs hit was definitely subtler than originally expected and there were certainly many exceptions to the general trend. For example, the stadium with the lowest altitude (New York Yankees) recorded more home runs than the stadium with the highest altitude (Colorado Rockies). There are also many stadiums with high altitudes that reported less home runs than average and many stadiums with low altitudes that reported more home runs than average. Because altitude did not seem to create a huge difference in the number of home runs recorded, we are unable to use this bar chart to reach any conclusions. However, this only shows that there are other statistics that have to be accounted for before determining the effects of altitude on offensive statistics.

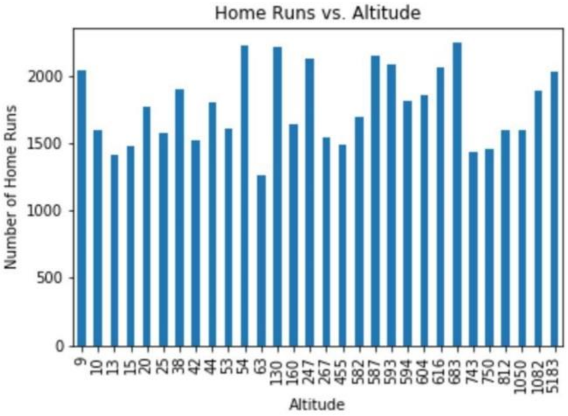


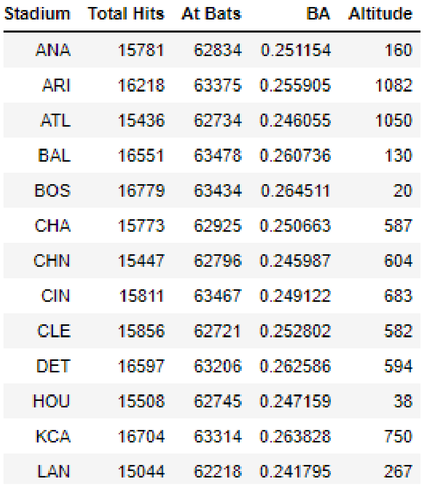
Figure 4: A Bar chart showing number of home runs hit at different altitudes

Finally, just to be sure, we put the home run data into a scatter plot to give us a better sense of any correlation. As can be seen above, there is positive. As we have said, this is just one stat, but so far it does not seem like altitude has much effect on offensive output in baseball.

9.2 KEY RESULTS

<u>Batting Average:</u> One statistic we focused on was batting average at each stadium. Batting Average is the total number of hits (singles, doubles, triples, home runs) divided by the

number of total at bats (strikeouts, flyouts, etc.). A sample of the results we got are in Figure 5. (This was because the table took up so much room)



Stadium	Total Hits	At Bats	BA	Altitude
ANA	15781	62834	0.251154	160
ARI	16218	63375	0.255905	1082
ATL	15436	62734	0.246055	1050
BAL	16551	63478	0.260736	130
BOS	16779	63434	0.264511	20
CHA	15773	62925	0.250663	587
CHN	15447	62796	0.245987	604
CIN	15811	63467	0.249122	683
CLE	15856	62721	0.252802	582
DET	16597	63206	0.262586	594
HOU	15508	62745	0.247159	38
KCA	16704	63314	0.263828	750
LAN	15044	62218	0.241795	267

Figure 5: A table that shows some of the teams stadium, total hits, at bats, batting average, and the altitude the games are played at.

Once we got this data we found that there was a correlation of .6 between altitude and batting average, which mean as the altitude gets higher batting average does increase, however we are not sure if there is enough evidence to say one causes the other. We then graphed a scatterplot with altitude on the x-axis and batting average on the y-axis to see if we could display the correlation. This is shown on Figure 6.

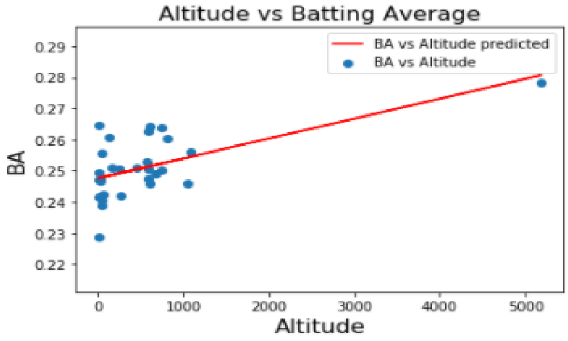


Figure 6: A table that shows each home teams stadium, total hits, at bats, batting average, and the altitude the games are played at.

With the above graph you can see that there is an outlier, which is COL. It is about 4000ft higher than the next closest stadium and has a higher batting average. The correlation decreased by a large amount going from a .6 to a .3. So we decided to graph the results again but without COL. This can be seen in Figure 7.

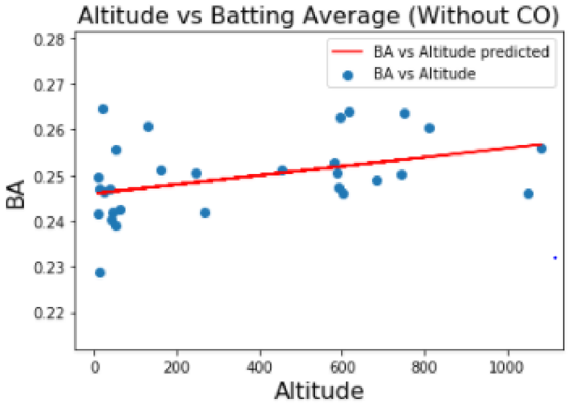


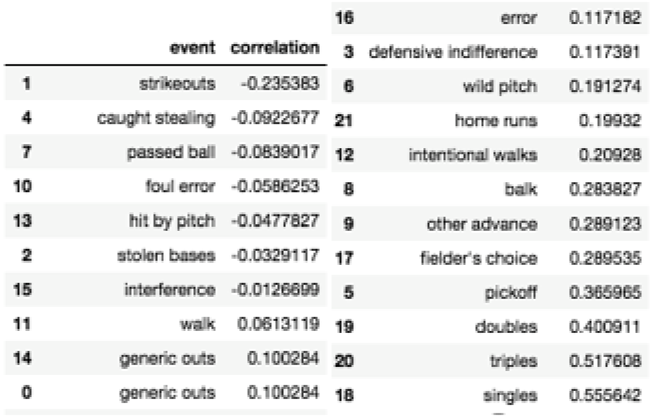
Figure 7: A graph showing altitude vs batting average without Colorado.

Once we removed the outlier is helped us conclude that batting average is not highly effected by altitude as you can see in Figure 7 the batting averages at stadiums are very much spread across the graph no matter the altitude.

Correlation of Events:

The first data mining technique that we implemented was iterating over all the events in our dataset to determine which events had the strongest correlation to altitude. Figure 8 includes some of the events/correlations that we considered to be of high important to game of baseball.

As we predicted, the statistics that involved batting had relatively high correlations to altitude. Moreover, it is evident that of these statistics, singles had the highest correlation with altitude - that is 0.555642. In order to understand the relationship between singles and altitude more thoroughly, we plotted these attributes against each other using both bar graphs and scatter plots. These diagrams are depicted in Figure 9 and 10.



	event	correlation
1	strikeouts	-0.235383
4	caught stealing	-0.0922677
7	passed ball	-0.0839017
10	foul error	-0.0586253
13	hit by pitch	-0.0477827
2	stolen bases	-0.0329117
15	interference	-0.0126699
11	walk	0.0613119
14	generic outs	0.100284
0	generic outs	0.100284
16	error	0.117182
3	defensive indifference	0.117391
6	wild pitch	0.191274
21	home runs	0.19932
12	intentional walks	0.20928
8	balk	0.283827
9	other advance	0.289123
17	fielder's choice	0.289535
5	pickoff	0.365965
19	doubles	0.400911
20	triples	0.517608
18	singles	0.555642

Figure 8: Correlation based on different events that occur during a baseball game and the altitude the game is played at.

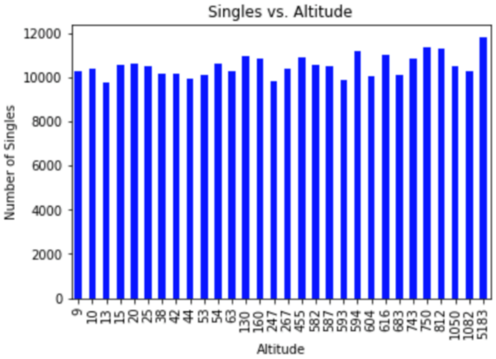


Figure 9: Bar graph showing the number of singles based on altitude

It is evident in the Singles vs. Altitude scatter plot, that the number of singles tends to increase as altitude increases. This relationship is illustrated by the least squares regression line (LSRL) found in the scatterplot. It is also important to note that the confidence interval (the light red zone in the scatter plot) also becomes larger as altitude increases. The widening of the confidence interval indicates that as altitude increases the confidence of our predicted line decreases. By fault, the confidence interval raises concern in regards to the credibility of the relationship between singles and altitude.

Another statistic that we found surprising was the correlation between home runs and altitude. We initially believed that home runs should have the strongest correlation to altitude, since the ball leaves the bat higher velocities at higher altitudes. However, after mining the data, it was evident that the number of home runs had a weak correlation with altitude - this correlation was 0.19932. Again, in order to thoroughly understand the relationship between home runs and altitude we plotted them side by side. These plots are shown in Figures 11 and 12.

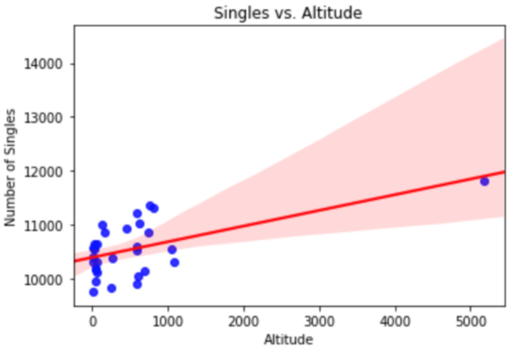


Figure 10: Scatterplot showing singles vs altitude with a line of regression and the red shaded areas are the confidence interval.

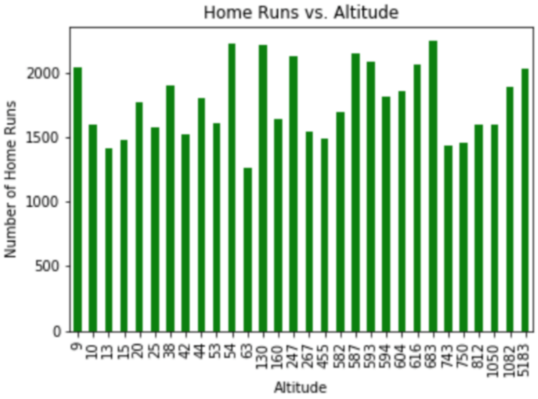


Figure 11: Bar graph showing the number of home runs based on altitude.

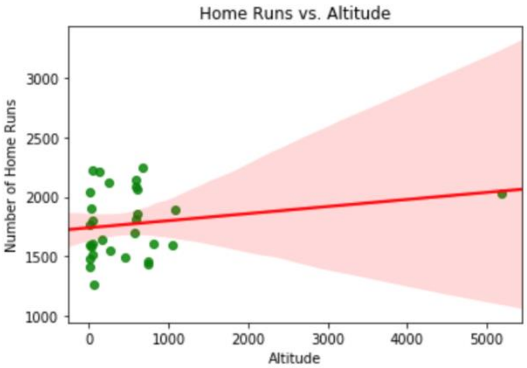


Figure 12: Scatterplot showing home runs vs altitude with a line of regression and the red shaded areas are the confidence interval.

Figures 11 and 12 illustrate that the correlation between the number of home runs and altitude is not as strong as we had initially anticipated. Also, the confidence interval in the Home Runs vs. Altitude scatterplot follows the same trend as the Singles vs. Altitude scatterplot, that is the confidence interval increases as the altitude increases. Consequently, these diagrams indicate that the effect that altitude has on home runs is miniscule.

Runs Per Nine:

Another area we decided to explore was the total runs scored at each stadium in the league. If altitude does have an impact on the offensive performance of players, then it would make sense to see more runs scored at stadiums with higher elevations. One big thing we had to account for with this statistic was the fact that not every stadium has had the same number of innings played. Extra inning games occur all the time in baseball. Some of these games only need one extra inning, and some need ten. Because of this, we cannot simply compare all runs scored in each stadium, as this would not be a true, normalized comparison. To account for this, we used a statistic called runs per 9 innings. To calculate this stat, you multiply the total number of runs scored by 9, then divide that quantity by the total number of innings played. Doing this will give an average of runs scored per 9 innings at each stadium, and we know that it will be normalized.

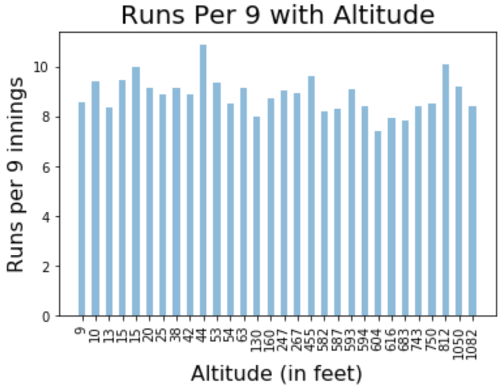


Figure 12: Bar graph showing Runs per nine vs altitude

Figure 12 is a graph that we were able to create from our initial results. From first glance, any positive correlation with altitude does not look promising. To look a little more in depth though, we also put the data into a scatter plot (Figure 13)

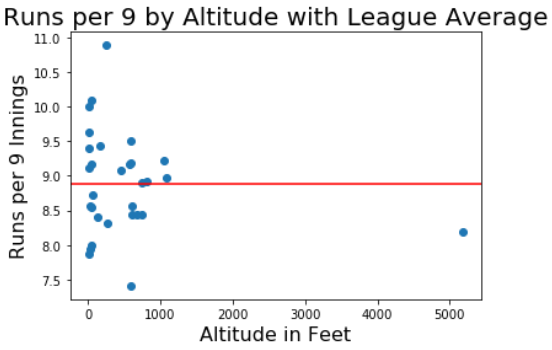


Figure 13: Bar graph showing Runs per nine vs altitude

The plot shown in Figure 13 is a scatter plot with the runs per 9 data. The red line represents the league average runs scored per 9 innings, which was about 8.88 runs. From this you can see that there is a proportional number of teams above and below the average at both high and low stadiums. Again, this is not promising for a positive correlation.

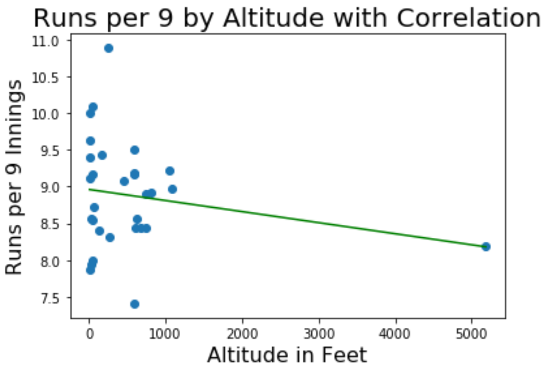


Figure 14: Scatterplot showing Runs per nine vs altitude with a linear regression

Finally, we plotted the runs per 9 innings data (Figure 14), and drew a least squares regression line. As can be seen above, there was actually a slight negative correlation between runs scored per 9 innings and altitude. While this is not a very significant correlation, it proves to us that the total number of runs scored at a stadium is not affected by the elevation of that stadium.

Home Run Fraction:

The statistic that altitude should have the most effect on is home runs. Based on our research, the largest effect that altitude has on the game of baseball is how far a ball travels in the air. If it is at a higher altitude, meaning thinner air, then the ball will travel further than a ball hit at lower altitude. Thus, if a ball is hit hard in the air, then it should have a better chance of being a home run at a higher elevation. After some research, we found that the best way to track this was with a stat called home run fraction. This is the number of runs scored via home run as a fraction of total runs scored in each park.

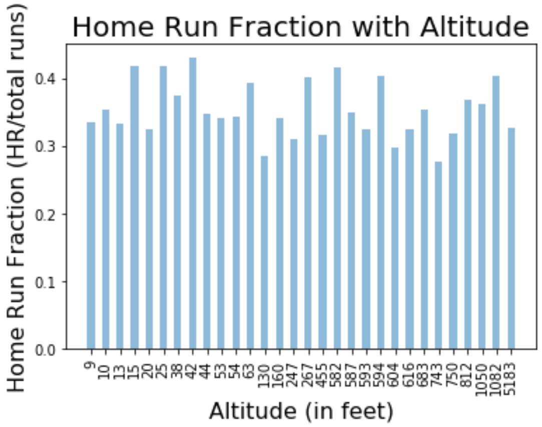


Figure 15: Bar Graph showing Home Run Fraction vs Altitude

Figure 15 shows the results of home run fraction vs. altitude. This graph shows a lot of disparity, but at first glance it does not seem to show a correlation between the two variables.

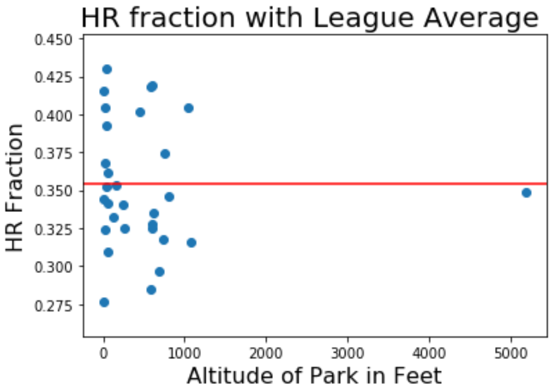


Figure 16: Scatterplot showing HR fraction vs altitude and the league average.

When the data is put into a scatter plot with the league average, you can see that there are significantly more teams below the league average than above it (including the Rockies), but these teams are equally distributed between low and high altitudes. The league average HR fraction was about 0.35. One interesting thing that can be taken away from this scatter plot is that the max and the min are both very close in altitude. This fact in itself almost guarantees that there is no relationship between altitude and HR fraction. This was proved in the below scatter plot, in which we added the least squares regression line.

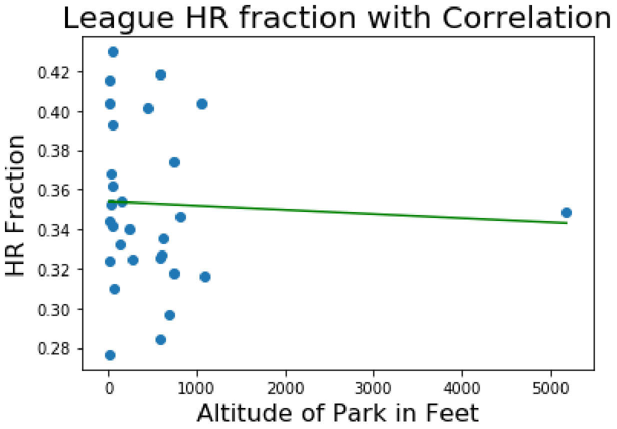


Figure 17: Scatterplot showing HR fraction vs altitude with line of regression

Note: Looking at this data, it is very apparent that we have one huge outlier when it comes to altitude, and that is Coors Field. Coors Field sits at just about a mile high, while the next highest stadium is Chase Field in Arizona, which is at 1,082 feet. The stadium altitudes up to Chase Field are evenly distributed, it is just the mile high elevation that sticks out. This really influences our graphs and visualization as well. As just one example, the below graph is the correlation of HR fraction and altitude with the Rockies' data taken out.

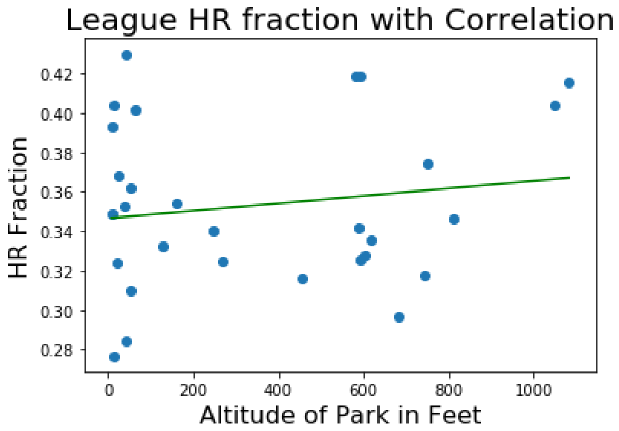


Figure 18: Scatterplot showing HR fraction vs altitude with a linear regression without COL

Notice that this is a huge difference! Without the Rockies' data, the correlation completely flipped from a negative correlation to a positive correlation. This at least raises the question of what the correlation would look like if the altitudes were somewhat uniformly distributed. Unfortunately, we cannot explore this option, at least for data in the MLB, as there are no more stadiums to explore. One thing we could look at in the future is data at different stadiums across the country, not necessarily from the MLB. By doing this, we would hopefully get a better distribution of altitudes, and maybe we could get a whole different result.

<u>Individial Player HomeRuns:</u>

For most of our work thus far, we have analyzed the statistics recorded out of the collective efforts of entire teams at specific stadiums. Another way we could try to find a correlation between altitude and performance would be to isolate the statistics of individual players to see if they performed better when playing in stadiums with higher altitudes. In order to accomplish this, we decided to analyze the top five players with the most homeruns between the years 2005 and 2015. These players were Albert Pujols, David Ortiz, Miguel Cabrera, Ryan Howard, and Adam Dunn. For these players, we were most interested in the number of homeruns they hit at each stadium because this is the statistic that altitude should have the greatest impact on. Unfortunately we were unable to find a big correlation between the altitude and the number of home runs scored for any of these players.

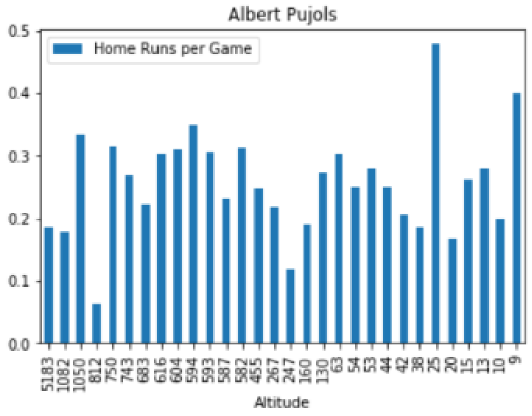


Figure 19: Bar graph showing the homeruns per game statistic for Albert Pujols when playing at different altitudes.

Figure 19 above shows the results for Albert Pujols when calculating the average number of homeruns he scored at each altitude. Similarly to the other players we analyzed, we were unable to find any reason to claim that playing at higher altitudes allowed him to score more homeruns. When applying the data to a scatter plot and applying a least squares regression line, we were unable to make any new conclusions. Adam Dunn and Ryan Howard were shown to play better with altitude, although not significantly. As shown in Figure 20, the positive correlation between altitude and homeruns per game for Adam Dunn is certainly a statistic we were proud to have discovered. However, the other three players we analyzed were actually shown to play worse at higher altitudes. For this reason, we were unable to declare that altitude provides a direct correlation with player performance.

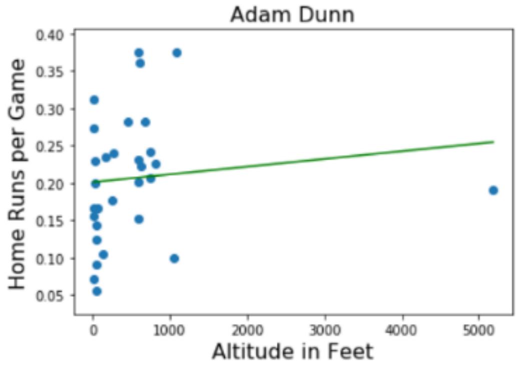


Figure 20: Scatter plot showing the home runs per game statistics for each altitude that Adam Dunn played with a least squares linear regression line.

10 APPLICATIONS OF RESULTS

Coors field has the reputation for being a great hitters park because of its altitude. After our investigation though, we have shown that altitude does not have a significant effect on offensive statistics. This could be utilized by scouting reports across the league. Because we were not able to find a correlation, then teams will not need to spend time trying to normalize stats that occurred at high altitudes. Additionally, if someone is looking to create a new team, our results show that placing this new team in a place with high elevation will not necessarily result in them becoming a better team.

As mentioned in the previous note, one application gained from this study would be to look into different stadiums at different elevations. As we saw, Coors Field, the one outlier in terms of altitude, really threw off some of our measurements. So if we were able to get data on stadiums at all different elevations, then maybe we would see more of a correlation between offensive production and altitude.