Park Factor in Relation to Everyday Baseball Statistics

In the world of baseball, there exists a multitude of factors that can influence a team's ability to score runs. Some of the more commonly accepted and heavily utilized factors that sports statisticians use to predict the outcomes of the games are earned run average (ERA), slugging percentage (SLG), runs scored, and total outfield area. For the purpose of our project we would like to explore a relatively new statistic in baseball, Park Factor. Although there currently exists a variety of methods to quantify Park Factor, we chose to use ESPN's version of the statistic: ((Runs Scored at Home+Runs Allowed at Home)/(# of Home Games))/((Runs Scored on Road+Runs Allowed at Home)/(# of Away Games)). As this simplifies to a ratio that compares the total runs scored at home with the total runs scored on the road, a result greater than 1.0 thus favors the hitter, while a result lower than 1.0 favors the pitcher. What makes Park Factor such a unique statistic is the reality that no other major sport employs playing fields with glaringly varying dimensions. In essence, the dimensions of each field vary in size and total area (excluding the inner diamond). In the outfield, the landscape of the field and architecture of the stadium can be vastly unique, where some of the outfield walls can scale 40 feet to a mere 3 feet in height. Other additional factors that can severely affect the outcomes of games include type of field (grass or turf), altitude, and typical weather patterns in the specified geographical area. Thus, it is possible that certain stadiums can be more advantageous than others in either hitting or pitching, and can consequently result in optimizing certain types of offensive or pitching performances, which in turn can significantly alter the outcome of any given game. Ultimately, our objective is to determine if there is a significant relationship between Park Factor and other commonly used baseball statistics such as ERA, OPS, SLG, runs scored and total outfield area, and whether or not it can translate into an alteration of a team's winning percentage (at home). By understanding the relationship between park factor and the aforementioned common baseball statistics, we can theoretically alter strategies based on playing field, assess a team's ability to adapt to certain parks, predict a team's average number of runs scored, and ultimately their overall performance over the course of a season.

With Park Factor being such a relatively new statistic, we chose to implement a data set spanning over the course of the past nine baseball seasons, with each season being the sum of 162 games played by 30 individual teams. Our data includes each baseball team, earned run average (ERA), slugging percentage (SLG), runs scored at home (RSH), runs scored away (RSA), runs allowed at home (RAH), runs allowed away (RAA), on base percentage(OBP), and the total outfield area of each playing field. As we expected a variety of data cleaning, we chose to calculate the Park Factor ourselves after a process of data vetting and cleaning. One such example of the data cleaning involves the notion that the span of our data set incorporates a variety of stadiums that no longer in use by the current 30 teams in professional baseball. As Park Factor is a statistic that relies solely on the effect that a certain playing field has on the performances of players, we made the decision to eliminate data stemming from parks that no longer exist. Another element of our data cleaning process involved the investigation of the true meaning of "home" and "away" games, as a select few number of games were played at alternate stadiums not listed on the team's record, with the home/away statistic leaving no room for an asterisk. Due to this type inconsistency, we chose to also eliminate these data from our overall set. With this cleaning process in mind, we aimed to eradicate any data that would skew our results in a manner that would not be truly representative of the realistic state of baseball, ultimately with the goal of gaining as a pure of an understanding of Park Factor as we can.

Initially, we were under the perception that Park Factor and ERA would show the strongest correlation amongst the aforementioned variables. From a purely baseball perspective, it seemed logical to believe that, Park Factor, a statistic that analyzes whether or not a given stadium plays more friendly towards either hitters or pitchers, and the team's overall ERA, a measure that quantifies the strength of a pitching staff by distributing their earned runs over a factor of nine innings, would embody some sort of quantifiable linear relationship. More specifically, we postulated that a pitching staff with a higher ERA would correlate with a Park Factor of over 1.0; similarly, a pitching staff with a lower ERA would correlate with a Park Factor of less than 1.0. Nevertheless, to our surprise, this was not the case. We came to the realization that using the team's overall ERA for all 162 regular season games over the course of the season was not the best method to use. With this, there was bound to be an inaccurate representation of how correlated these two were since the overall ERA includes all away games as well, and each individual team plays at many different away stadiums over the course of those 81 games played away from home. In order to get a large enough sample size to analyze how pitching staffs would perform in other stadiums, we decided that it would be necessary to collect data from a much larger range; however, this would be somewhat unrealistic due to the variation in teams and the turnover of new stadiums.

Ultimately, we came across the solution to create an ERA Ratio, which would compare the ERA of teams while playing at their home parks with the corresponding ERA of their road games. By dividing these two variables, we could finally see whether or not each individual team had a pitching staff that performed more favorably on the road or at home, or was simply neutral and had fairly consistent performance over the course of the season, regardless of the playing field. While a strong correlation between Park Factor and ERA Ratio seemed imminent, the resulting R squared value of 0.2855 shows that our predictor ERA Ratio explains 28.55% of our the variation in our Park Factor. And looking at the slope of our linear model, for every 1 point increase in our ERA ratio, the Park Factor increases by 0.4655.

After examining the pitching side of the game, focusing more on the offensive end of baseball, it seemed as if a team's OBP (the percentage of at bats where the player would get on base via a walk or a hit) would also be an accurate predictor of park factor. The expectation was that this would not necessarily be truly indicative of what a team's park factor was. The thought initially was that OBP would not be the most useful statistic when predicting park factor, since OBP is likely more representative of the team's talent and Park Factor's purpose is only analyzing the impact of the stadium on each game. Depending on how management has designed a team's talent, will adjust the value of OBP quite significantly. For example, if a baseball team is more pitching oriented than batting, it is likely that it will have a lower team OBP than that of a team that has a roster assembled to be more potent offensively. Even with this being the case, the general consensus was that it was expected to have a weaker correlation coefficient than when ERA was as a predictor for park factor. Surprisingly enough, it actually had a slightly stronger relationship than ERA and Park Factor. Our R squared value shows us that our predictor of On Base Percentage explains 41.67% of the variation in our Park Factor. And looking at the slope of our linear model, for every 1 point increase in our OBP ratio, the Park Factor increases by 1.2895.

Following our intuition, we wanted to consider the relationship between slugging (SLG), a measure of the power of the hitter, and Park Factor. We believed this would be an accurate variable for predicting the response because logically, a greater power percentage produced by particular team would in return have a direct correlation to a greater number of total runs, thus directly increasing the overall ratio quantified by the Park Factor statistic. Thus from our hypothesis, we believed there would be a

stronger correlation between slugging and Park Factor. After running and examining our code, we found that slugging had a positive linear relation with Park Factor with a cluster of data between 0.9 to 1.2 in the x-axis and 0.8 to 1.2 in the y-axis. Additionally, we found that slugging explained 61.01 percent of the variation of Park Factor, the highest correlation thus far. Then, by observing the residual errors in the residual plots, we established that the assumptions were met for each of the plots, with constant variance in the x-axis and the y-axis, normality in the distribution and no bad leverage points, thus supporting the strength of our correlation coefficient. Together with this data, we were able to conclude that there was a statistical significance between slugging percentage and Park Factor.

Since the total outfield area varies at every stadium, it would seem as if the outfield area would have a significant effect on the value of park factor. Often it is common in most sports that the dimensions of the playing field all would be identical, yet in baseball this is not really the case. There is no standard dimensions that a stadium has to follow when designing the outfields. Outfield area should be an important factor for a team's performance. If you consider a team that has the second largest outfield area in Major League Baseball, the Colorado Rockies (Coors Field), who have also boasted a playing field with the most favorable Park Factor to hitters during 8 of the past 9 seasons, this might cause a variety of questions to arise. One would come to the assumption that if a team had larger outfield area, naturally the more difficult it would be for batters to hit balls outside the park (which is noted by a Home Run) and that this would typically tend to favor the pitcher. In contrast, if a team had the smallest overall outfield area, one would tend to think that it would cause more balls to leave the confines of the playing field, thus resulting in a much higher scoring game. However, if you examine the aforementioned Coors Field, which as previously mentioned, has the second largest outfield area in baseball, also has an unparalleled Park Factor ratio. While this could be an anomaly of sorts, our results indicated that outfield area was not a significant predictor of Park Factor, which, once again, disproves our previous assertions. Our R squared value shows us that our predictor of Outfield Area explains only 4.552% of the variation in our Park Factor. After examining at the slope of our linear model, for every 1 point increase in our OBP ratio, the Park Factor increases by 0.008296, a very insignificant increase.

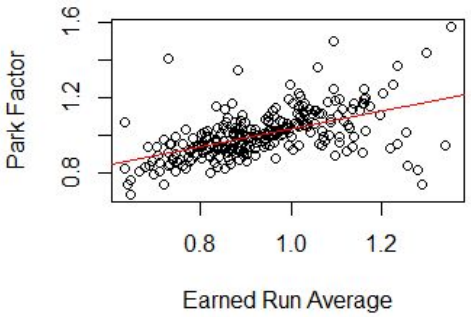
Lastly, we wanted to take our project a step further by determining if home winning percentage could be correlated with Park Factor. We wanted to test the assumption that the team's with greater winning percentages at their home games had more positive ratios greater than one; the ratio that would be advantageous towards hitters. We believed there should be a significant relationship because teams would have more experience when practicing and playing at their home stadium, through trial and error. Moreso they would have more knowledge of the uniqueness of their stadium's architecture and field. To our surprise, we found that there was actually an array of scattered data, suggesting a weak linear regression with Park Factor. Also, the residual errors in our residual plots did not meet the correct assumption for each plot (Residual vs Fitted, QQ Plot, Scale Location, and Residuals vs Leverage). The cluster of data in residuals and outliers tell us that our r-squared value of 0.005684, the lowest in comparison to the other predictors, may be inaccurate as well. After considering our results from the regression and scatter plot we came to the realization that winning percentages cannot be solely correlated to Park Factor because of the large range of other factors that influence the results of each home game. Such factors include, the random variation of the performances of each individual pitcher and hitters, climate/weather, altitude, skewed right or left handed power hitters and player injuries which greatly affect each game and season, which is reinforced by the random data points when plotted. We concluded

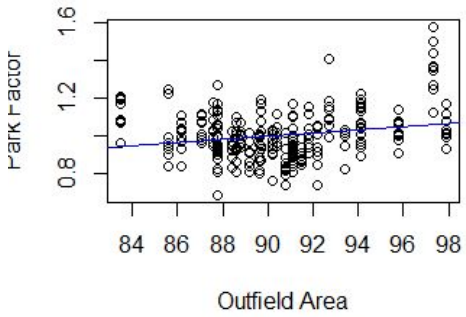
that because of the multitude of factors within a single baseball game, home-winning percentage would have the weakest association with park factor.

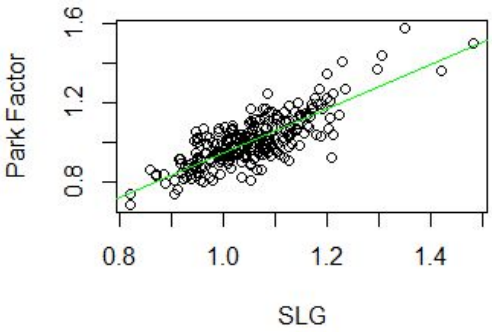
For our multiple linear regression, we want to prove and defend our results showing that our ERA ratio, On Base Percentage Ratio, and Slugging Percentage Ratio are indeed significant predictors for our response variable, Park Factor. Therefore, we create a full model of multiple predictors with Park Factor as the response variable. When we run the code, our full model shows three significant predictors. Looking at the p-values of we notice that our p-values for ERA ratio, OBP ratio, and SLG ratio are all approximately 0. Because our p-values are less that 0.05, because we are doing a T-test, we can say that these three predictors are significant predictors in our MLR. To continue and prove that these three are the only significant predictors in our data, we create a reduced model, deleting Outfield Area and Home Winning Percentages as predictors. After deleting, these predictors we compare the adjusted R^2 value from the full model to the reduced model. We notice a very insignificant difference, from 0.7528 in our full model to 0.7509 in our reduced model. And looking at our reduced model, we notice and can conclude again that ERA ratio, OBP ratio, and SLG ratio are significant predictors for our response variable, Park Factor.

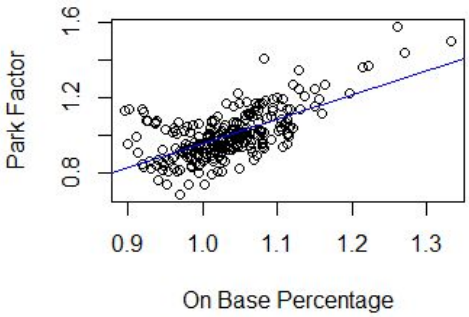
When further contemplating the relationship between Park Factor and the previously analyzed statistics, we proceeded to reexamine its correlation with ERA ratio, but on a more constrained level. Rather than utilizing the nine-year span we implemented for our data, we chose to use a much more limited two-year span with data from each game from the years 2015-2016. With the understanding that baseball rosters undergo severe personnel turnover constantly, and especially within the scope of a nine-year span, we postulated that a smaller data set might provide us with a more accurate (although somewhat limited) data set. Our R squared value shows us that our predictor of ERA ratio with 2 years of data explains a high 81.04% of the variation in our Park Factor. And looking at the slope of our linear model, for every 1 point increase in our ERA ratio, there is a 0.8653 increase in our Park Factor.

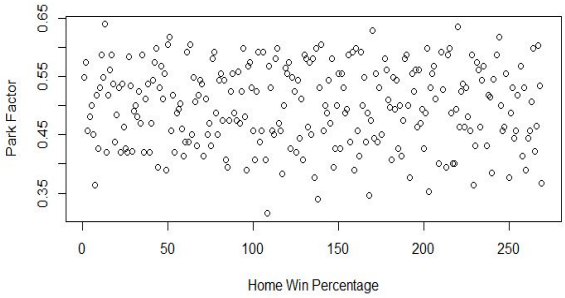
Unfortunately, our analysis naturally had some flaws that we simply did not have much control over. More than any other sport, baseball tends to have the perception that individual performance varies significantly on a year to year basis. So with this often random variation in performance, it could potentially cause our analysis to be inaccurate. In addition, the MLB is made up of two different conferences, the American League (AL) and the National League (NL). In the NL, the pitcher bats in the 9th spot in the batting lineup. However, in the AL, the pitcher does not bat at all. In fact they have a player that is referred to as the “Designated Hitter” and this player takes the spot of the pitcher in the lineup. While there definitely are offensive benefits to this(since pitchers generally considered to be not the greatest hitters), it is easy to see that it would be ideal to separate the NL from the AL. The most significant flaw with separating the two leagues, is that depending on who they play, whether or not they use a designated hitter often changes. For example, in interleague play(a team from the NL plays the AL), depending on who is the home team and what league they are from determines whether or not the teams play with a designated hitter. So in order for this to be useful we would need to discover which away games a team played used a designated hitter and which away games did not. We found it beneficial to just analyze the two leagues as a whole and assume that this did not play a large factor. So the differences between the two leagues could be a limitation of our analysis. There are many other things that could hinder how accurate our analysis is, such as differences in weather and climate, altitude, constant roster turnover from year to year, and there also many other limitations, but we considered these to be the most important to identify.












```
> fullmultiplemlb <- lm(MLB_2012$PF ~ MLB_2012$ERAratio + MLB_2012$OBPratio + MLB_2012$SLGratio + MLB_20
12$OF.Area + MLB_2012$hwp)
> summary(fullmultiplemlb)

Call:
lm(formula = MLB_2012$PF ~ MLB_2012$ERAratio + MLB_2012$OBPratio +
    MLB_2012$SLGratio + MLB_2012$OF.Area + MLB_2012$hwp)

Residuals:
      Min        1Q    Median        3Q
-0.221697 -0.034114  0.002358  0.032961
      Max
 0.296130

Coefficients:
                Estimate Std. Error
(Intercept)     -0.499170   0.121933
MLB_2012$ERAratio  0.257786   0.028496
MLB_2012$OBPratio  0.295642   0.083154
MLB_2012$SLGratio  0.844545   0.057629
MLB_2012$OF.Area   0.001302   0.001230
MLB_2012$hwp      -0.094498   0.058290
                t value Pr(>|t|)
(Intercept)      -4.094 5.69e-05 ***
MLB_2012$ERAratio   9.046  < 2e-16 ***
MLB_2012$OBPratio   3.555 0.000449 ***
MLB_2012$SLGratio  14.655  < 2e-16 ***
MLB_2012$OF.Area    1.059 0.290604
MLB_2012$hwp       -1.621 0.106216
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06396 on 256 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.7575,    Adjusted R-squared:  0.7528
F-statistic: 159.9 on 5 and 256 DF,  p-value: < 2.2e-16
```