**Introduction**

My individual project for this course is going to focus on baseball analytics. Baseball statistics have a lot of in-depth data to describe their offensive and defensive stats. I am investigating the Chicago Cubs pitching and offensive data to ideally answer the question of what leads to wins/losses. The reason I am choosing to investigate correlation coefficients to wins and losses is improvements could then be brought to the organization to show what we should continue to strive in as it leads to wins, or vice versa when a stat is shown to lead to losses training can be done additionally in hopes to improve their record.

My data was retrieved from https://www.baseball-reference.com which has plenty of sports data even outside of baseball. Luckily this site is very easy to navigate and get a look at what the raw data will look like once downloaded. There is a super easy export option to get the data in an excel file, the only odd part is it saves as an old excel file. This makes working on data in excel more tedious and slow performance wise, luckily copying and pasting the data in a new excel file everything transferred over correctly.

The pitching data has 25 columns of data while the offensive data has 27. Each data set lists the years of the Cubs organization (1876-2022) with a lot of stats. All the rows are numeric values and consistent with rounding. The website where the data was obtained has formulas for the more advanced statistics that are not self-explanatory. The data is mostly clean and not much work was needed to get into an analysis ready state, however some steps were taken.

Firstly, I got rid of the column “LG” as this just lists what league the organization was a part of that particular year. There could be statistical significance in what league they were in, even if there is the MLB organization handles league management. This stat would be pointless to include for the results I desire to obtain. Secondly, I added a column “W/L Ratio” (win loss ratio), each year contains the wins and losses for the appropriate season so it’s easy to calculate and allows for later analysis to find significance between wins, losses, and the W/L Ratio. Lastly, the offensive statistics have a glaring problem, the columns “SB” (stolen bases) and “CS” (caught stealing) have numerous rows with null values, these are the only columns with incomplete data. Therefore, these columns will be removed for the sake of consistency.

I had wished to compare the average values from the data provided to what I would find when working with the entire data set, however as analysis has been progressing, I am realizing for the output I have envisioned, I would have too many visuals and statistics. I wouldn’t want to include this and leave more short-end versions of all the information gathered. I was not expecting the raw data to provide me enough information. What I did in place of this is include frequencies / percentages in the excel spreadsheet to get even more information. I included a few columns, and the abbreviations are consistent, such as “HPG” which is “hits per game” and “DPPG” which is double plays per game. Obviously as an analyst working on this in a presentation setting this would be explained easily if it came up.

**Beginning Analysis**

Getting started with the analysis I quickly ran into a problem; I had wanted to create a heat map to show the statistical correlation between variables. The nice part of this map is since we are already calculating correlations, if the organization wanted to know if there were any other significant correlations, we could easily point that out with a visual heat map or all correlation values. However, since this is a data set with a lot of variables, a heat map of the entire data set is illegible. To work around this, I made separate worksheets in excel approximately 10-11 columns to make heat maps legible, while making sure I didn’t have too many visuals. I was able to get this to be four heat maps for offense, and another four for defense. The reason the columns are consistent is not only for aesthetic purposes looking at Excel or data frames, but this was also done so all the heat maps are the same size, or very close. I had to make sure I could get everything to a 10x10 or a 11x11 which worked out quite nicely.

This means I have very repetitive code in python for each worksheet but found it easier to work in excel for the sheets rather than make new data frames in python. To demonstrate that I can use python strictly for sorting data, I have four worksheets in excel for both offense and defense, the sheets: offense2 and defense2 are in excel for consistency purposes. I can then read the excel sheet and select the desired columns that would be equivalent to having a separate worksheet with just those columns. I then realized I needed to define more clearly what I was looking for, which is that I am looking for the top 5 correlations compared to 3 variables: wins, losses, and win loss ratio. We expect to see a lot of the same variables across these variables but is worth investing if something were to be different. We do not want wins, losses, and win loss ratio all in the same worksheet, otherwise we will see a high correlation between these variables and is useless information to have, this was tested and explained briefly later.

Luckily at this point in the analysis, I realized I would have repetitive code. Not only for the heat maps, but for finding the correlations desired, and the p-values. After finding the code that gave the output desired, I made the output accessible through functions. This helps with legibility and organization; as a bonus efficiency working through the appropriate steps. I made sure to add comments on every block of code, some of it is repetitive, but it shows the logical process working through this project and tweaking it for effectiveness. I had to make sure while doing all of this, variable and function names were logical, and that blocks of code were organized.

There was another problem that came up in the analysis, which is that W/L ratio comes up quite often in wins and losses and vice versa, this was unintended and wastes one of my top 5 correlation spots. Removing this will allow for a better output, not variables calculated from each other. To get around this, I had one worksheet without the win-loss ratio, and one with it that does not have the base number of wins and losses. This made a huge difference in the output and helps with clarity.

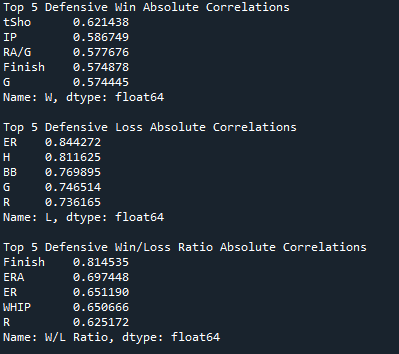
I knew I could begin finding the top 5 correlations between: wins, losses, and win-loss ratio. I must do this for both offense and defense and used sorting absolute values and iloc [1:6] for each of my outputs. The reason I must do 1:6 in my iloc function rather than 0:5 is the first value of 0:5 would be 1.000 and would be the variable we are comparing to; we need to negate this value and search for the next top 5 correlation values. We use absolute value to find the top 5 variables in correlation values, even though this will not tell us if the correlation is positive or negative, we can find that out later along with the p-values.

The results of the top 5 correlations are shown below the first set will be on my first analysis without frequencies calculated in excel, and the second set will be with the calculated values, this will show what difference the calculated values make on correlation findings:

**Top 5 defensive correlations Top 5 offensive correlations**

**without frequencies without frequencies**

**table 1 table 2**

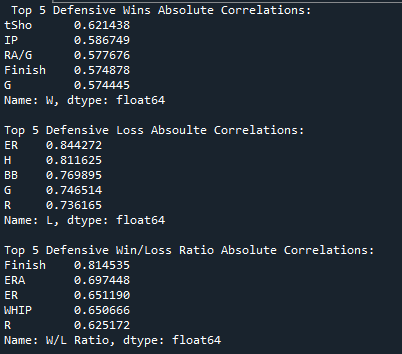
 Text

Description automatically generated

**Top 5 defensive correlations Top 5 offensive correlations**

**with frequencies with frequencies**

**table 3 table 4**

 **Text

Description automatically generated**

We must keep in mind that these correlation values seen are absolute values, meaning based on these numbers, we cannot determine if it is a positive or negative correlation in this context, this will be resolved once testing for significance begins. We could do steps in this stage to find out, but this will need to be done later anyways. It surprised me seeing these correlations, I have a background in sports so had an idea on what to expect to see in the output and have overlap; however, I was surprised to see many differences between wins, losses, and win-loss ratio and what impacts these stats. The fact we are seeing different variables in each correlation test means my methodology of testing these separately was a good idea. It was surprising to see significantly more recurrences in the offensive tests, whereas the pitching stats didn’t have near as many repeats.

I could have removed the column “Finish” as it is obvious that a higher placement should lead to more wins, lower placement leads to more losses. However, I chose to leave this in as if something has a higher correlation than this column it is worth noting despite this variable itself being self – explanatory. I could have also removed “G” which is games played, as there will only be more games played if there is a postseason for the Cubs, however I followed the same train of thought which is if we find correlations beyond these variables that should be at the top of our findings, it’s a good find and can emphasize how important our findings are then if we just took these out. This does create some “boring” variables and graphs but is important in the art of storytelling which will become relevant later in this paper with the visual representations of findings.

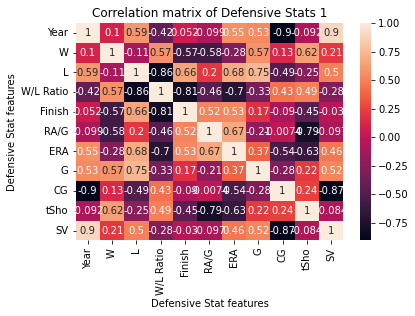
I was expecting higher correlation values, I was surprised to see in the top 5 correlation scores, we went as low as 0.39. This still gives us information that is useful with the questions I intended to answer. I still believe heat maps will be useful as even though we are looking specifically at what impacts wins, losses, and win-loss ratio, we can still from a brief look see what other instances we have a high or low correlation values. It will also be interesting to see data visualizations of the lower correlation scores as the graph may not be as representative as I am hoping, if this is the case data visualization may change as I would like the results to be strong, I would not want to present suggestions with sub-par significance.

When adding in the frequency calculations, it was interesting seeing no big changes in the top 5 correlation values, the only change was in our offensive win/loss ratio. It is alright that we didn’t have these values come up in our findings as much as we would have hoped, doing further calculations to enrich analysis is a strong fundamental to have and there could be scenarios outside of the analysis project where the findings could be useful to the organization.

Here we have the heat maps that have all the correlation values for every variable in the raw data and calculated frequencies. This is so if we were presenting this project to the team, we can have this information readily available and see the correlation values as a face value of significance, further testing could always be done for any findings. The heat map outputs are seen below:

**Defensive Stat Heat Maps**

**graph 1 graph2**

Chart, treemap chart

Description automatically generated

**graph 3 graph 4**

**Chart, treemap chart

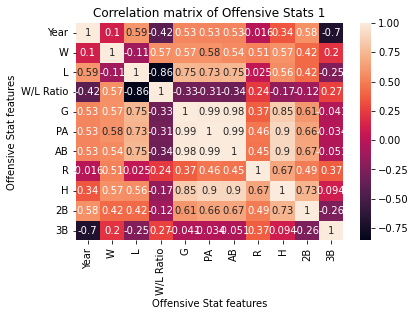
Description automatically generated**Chart, treemap chart

Description automatically generated

If this were done in a presentation setting, we can state how we can briefly see what strong or weak correlation values there are. I would point out that home runs, strikeouts, and save percentage have strong correlation values to the variables “Year” which means its either not consistent or has gotten strongly better/worse over time.

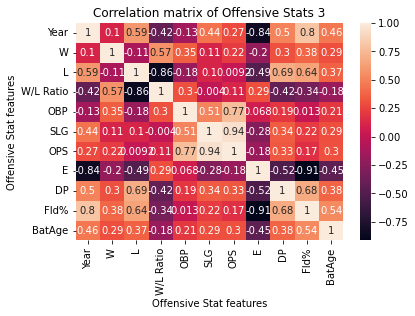
**Offensive Stat Heat Maps**

**graph 5 graph 6**

**Chart, treemap chart

Description automatically generated**

**graph 7 graph 8**

**Chart, treemap chart

Description automatically generated**

If this were done in a presentation setting, we can state how we can briefly see what strong or weak correlation values there are. Here we also see the home runs and strikeout have a correlation to the year variable, this could be mentioned our analysis team could see what has gotten better or worse over time. These are brief talking points that would be done in presentation not further findings in this project. It’s important to demonstrate why these heat maps were created.

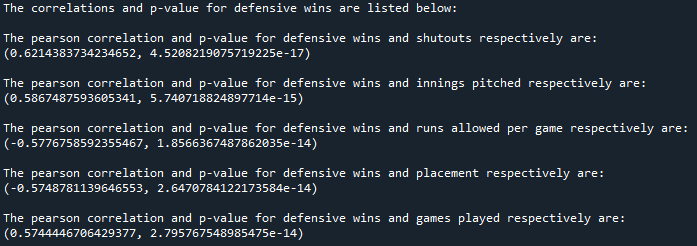
**Taking Analysis Further**

Now we can begin looking for statistical significance with the correlation values found, this will also be done in python, and we will have the display of the correlation (this time if it’s positive or negative) and the p-value. The reason this is done now rather than earlier, is we need to know the two variables we are comparing to utilize this command in python. The main reason we are using this command, is if we have a correlation with no significance, it’s like the concept of correlation doesn’t equate to causation. Since we get the returned p-value between the two variables tested, we are looking for a p-value of under 0.05 to say there is statistical significance.

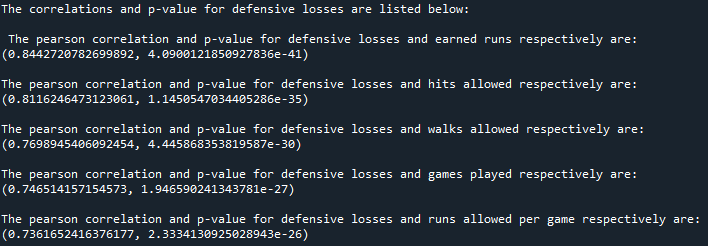
It is worth noting that we never want to look at statistical significance at face value as there are a lot of individual components making up the formula for the calculation. If we see high p values, we know that we are most likely looking at variables that do not have a strong relation to each other. From a large data set with a lot of variables and looking at the top 5 correlations, we expect to see good p-value scores. I have plans to look for more significance once I create visuals for the correlations, but this is a preliminary step to make sure the steps taken so far have been done correctly, and that we can proceed making an end result of this project.

The output received from the described analysis are as follows:

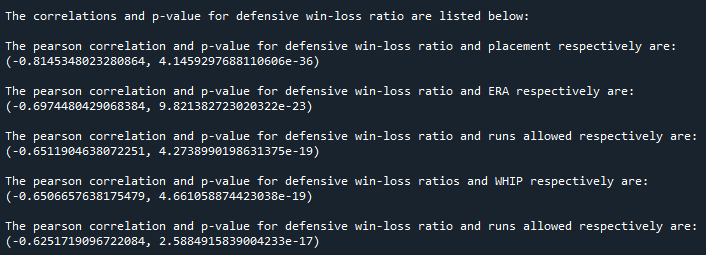
**Defensive Wins Correlations and P-Values (table 5)**



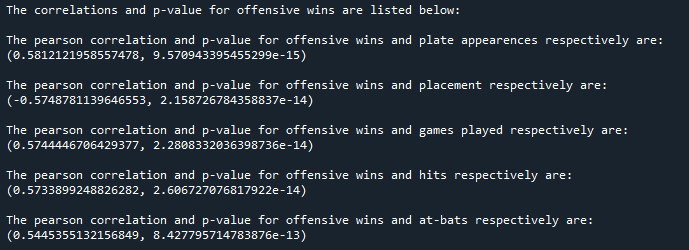
**Defensive Losses Correlations and P-Values (table 6)**



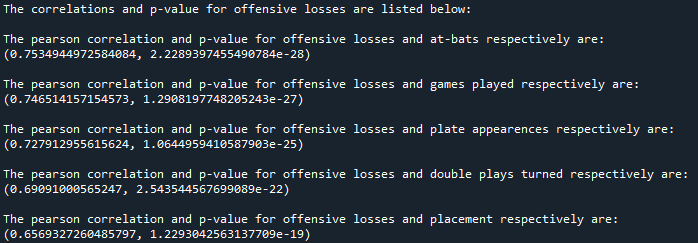
**Defensive Win-Loss Ratio Correlations and P-Values (table 7)**

****

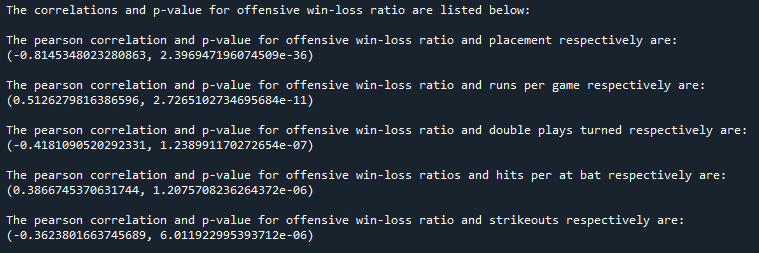
**Offensive Wins Correlations and P-Values (table 8)**

****

**Offensive Losses Correlations and P-Values (table 9)**



**Offensive Win-Loss Ratio Correlations and P-Values (table 10)**

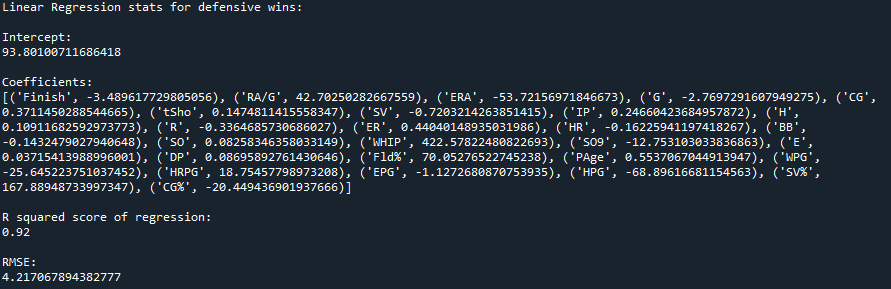
****

As we can see in the p-values returned, we would consider all variables as having statistical significance. We do not want to take this at face value as even a low correlation value of -0.36 (our lowest correlation value) is considered significant. Granted we cannot assume based on a correlation value if there is significance, but we can see with higher correlation values the number gets small with a huge negative exponent. The most probable reason to this happening is having a large data set, we have 147 rows of data in each variable tested. Having a large sample size, otherwise referred to as “N” can cause this to happen even if there is weak correlations and r-squared values.

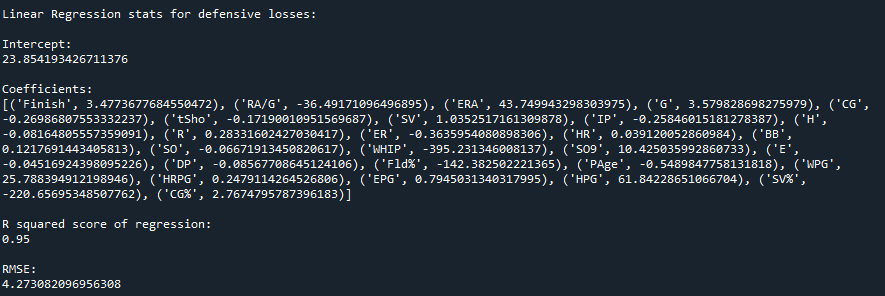
As visual representation is approached, this needs to be noted to look out for fits and trends to see how it looks compared to these findings of correlation values and p-values. We expect to see medium to strong correlations with low to decent fit. My plan is to work on this inside of excel as I know how to get the graphs, which will then be in the same spreadsheet as the data worked on. The main reason is to have the r squared values displayed on the graphs along with a linear equation. The equation is not necessary at all and isn’t going to be used for anything further in this project but showing that we can find an equation (calculated by a computer) for any 2 variables is a nice to have along with the main point. The r-squared value will give us an “accuracy” of our fit, we would ideally want to see a 0.9 or higher value to show a strong correlation, I am doubtful we will see a number that high as our correlation values are not as high as I had hoped for starting this project. We may see a few variables with a good r squared score, however I am doubtful based on findings thus far.

Before showing the scatter plot visuals to represent the correlation fit, we want to run a linear regression analysis on all variables to wins and losses. We cannot include win/loss ratio as this is a percentage and cannot be conducted with linear regression. This will tell us how strong our linear approach is, as this will also be used in predictive modeling later. Dealing with strictly numeric data and a lot of intrusive variables we expect to see high accuracy and results from this analysis step. On the following page you can see the stats from this step including coefficients, r squared score, and RMSE:

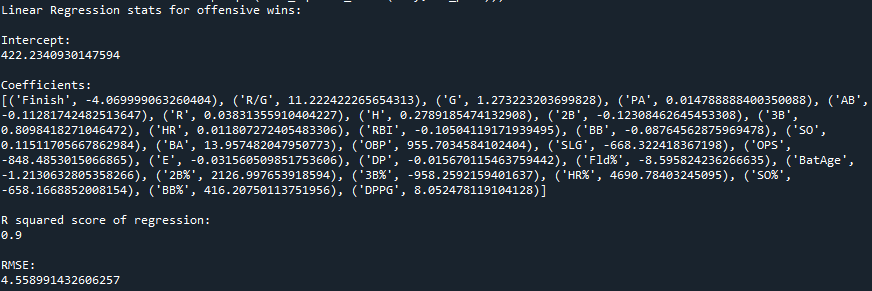
**Linear Regression for Defensive Wins (table 11)**



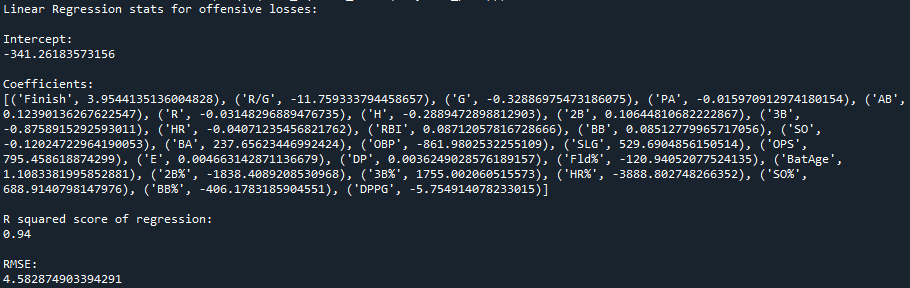
**Linear Regression for Defensive Losses (table 12)**



**Linear Regression for Offensive Wins (table 13)**

****

**Linear Regression for Offensive Losses (table 14)**

****

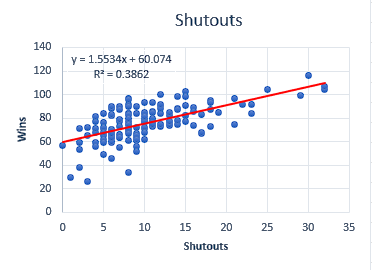
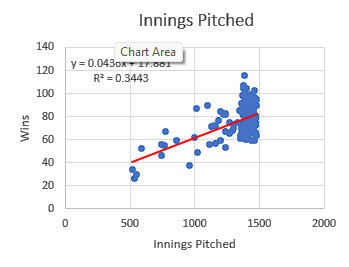
This shows that linear regression is a very strong approach for metrics in this project, as we can see in all of the regression test we have a 90% accuracy or higher. With a MSE value of less than 5. We have very strong coefficients and intercepts as well. Every value is what we would like to see in a strong regression result.

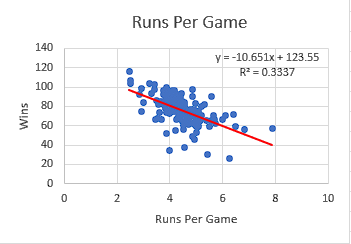
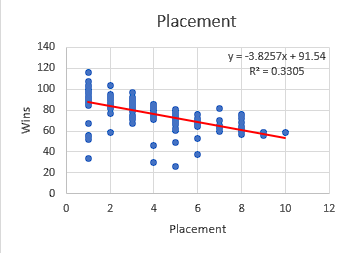
On the next few pages there will be a lot of graphs (30 to be exact). There wasn’t much actual work to make the graphs aside from repetitively grabbing the new variables, making sure formatting, visual presentation, font, centering all matched between all the graphs. There may not be a need for this many graphs in a paper-writing project, but feel in a presentation setting, using 6 slides (5 graphs each) it divides nicely to talk about each individual variable and the findings of the 5 variables researched. I will provide what I believe are the takeaway points after the graphs provided. Each graph has the r squared value, and the linear equation.

**Visual Representation**

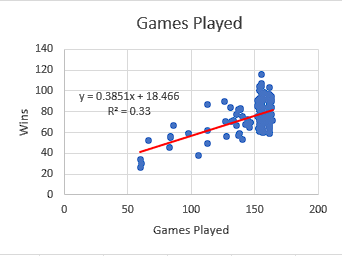
**Top 5 Defensive Wins Correlation Graphs**

**Defensive Win Scatter Plot 1 (graph 9) Defensive Win Scatter Plot 2 (graph 10)**

**Defensive Win Scatter Plot 3 (graph 11) Defensive Win Scatter Plot 4 (graph 12) **

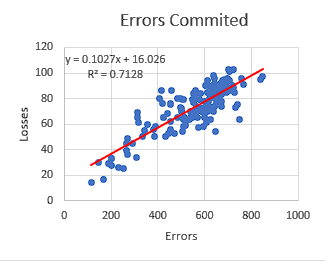
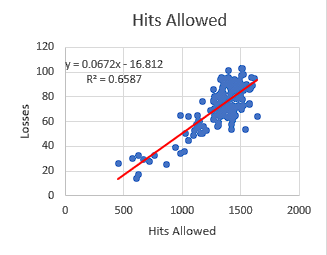
**Defensive Win Scatter Plot 5 (graph 13)**

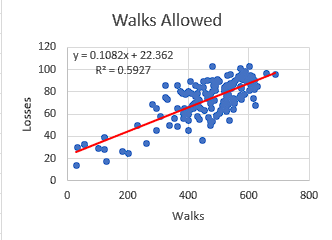
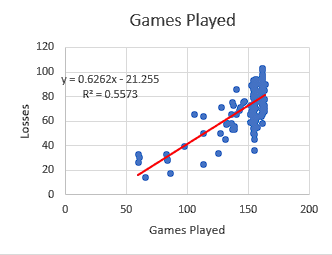


Take note of very low r-squared values showing that these variables have a poor correlation / fit of accuracy.

**Top 5 Defensive Losses Correlation Graphs**

**Defensive Loss Scatter Plot 1 (graph 14) Defensive Loss Scatter Plot 2 (graph 15)**

**Defensive Loss Scatter Plot 3 (graph 16) Defensive Loss Scatter Plot 4 (graph 17) **

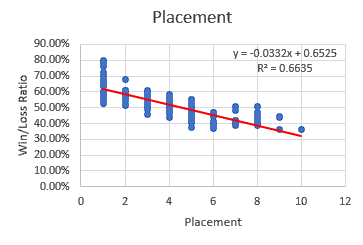
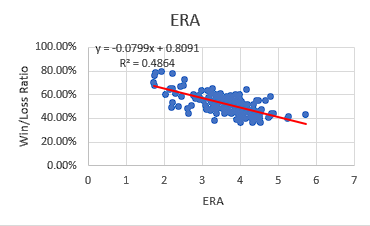
**Defensive Loss Scatter Plot 5 (graph 18)**

****

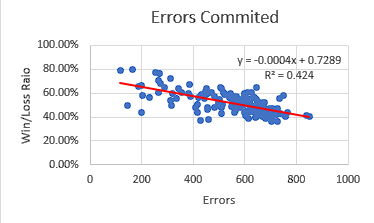
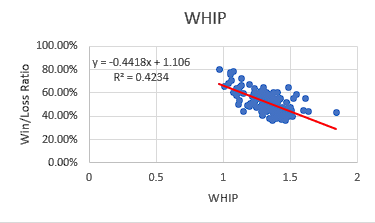
We see higher r-squared values showing these have a much stronger accuracy or correlation.

**Top 5 Defensive Win/Loss Ratio Correlation Graphs**

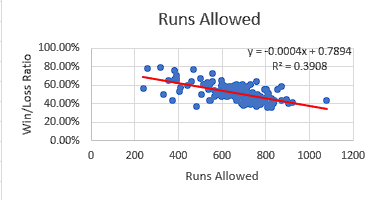
**Defensive Win/Loss Scatter Plot 1 (graph 19) Defensive Win/Loss Scatter Plot 2 (graph 20)**

**Defensive Win/Loss Scatter Plot 3 (graph 21) Defensive Win/Loss Scatter Plot 4 (graph 22)**

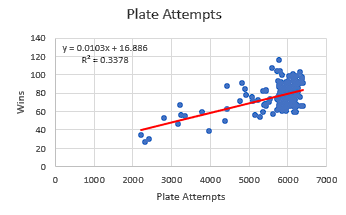
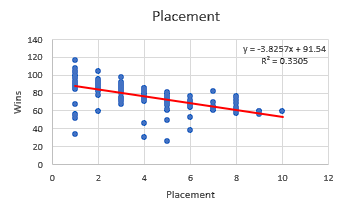
**Defensive Win/Loss Scatter Plot 5 (graph 23)**

****

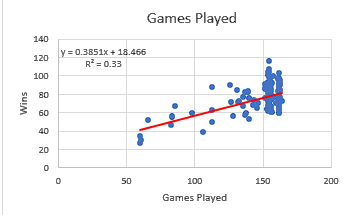
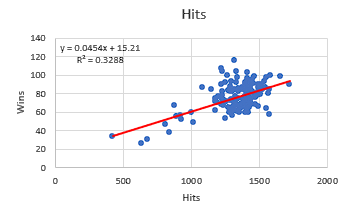
We see overall low r-squared scores showing these variables have relatively poor fit as well.

Note how placement has a significantly higher r-squared value compared to the other graphs.

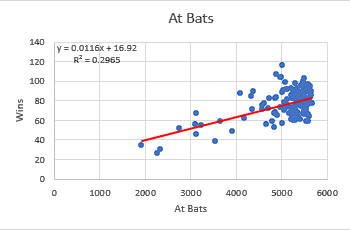
**Top 5 Offensive Wins Correlation Graphs**

**Offensive Win Scatter Plot 1 (graph 24) Offensive Win Scatter Plot 2 (graph 25)  **

**Offensive Win Scatter Plot 3 (graph 26) Offensive Win Scatter Plot 4 (graph 27)**

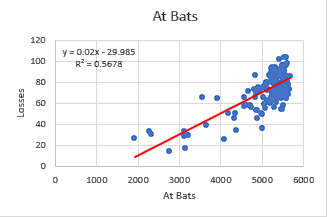
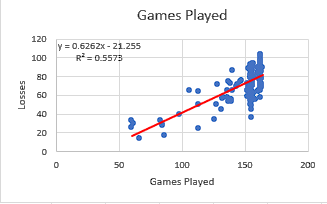
**Offensive Win Scatter Plot 5 (graph 28)**



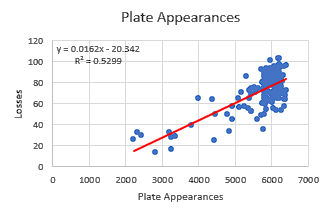
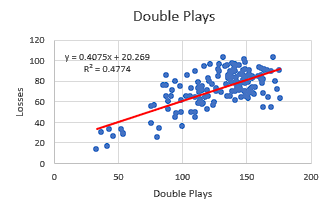
Note the low r-squared values similar to how we had low values in defensive wins as well, shows poor accuracy of correlation.

**Top 5 Offensive Losses Correlation Graphs**

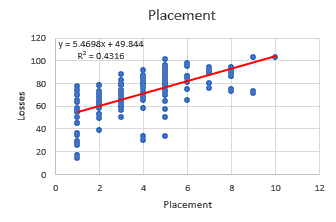
**Offensive Loss Scatter Plot 1 (graph 29) Offensive Loss Scatter Plot 2 (graph 30)**

**Offensive Loss Scatter Plot 3 (graph 31) Offensive Loss Scatter Plot 4 (graph 32)**

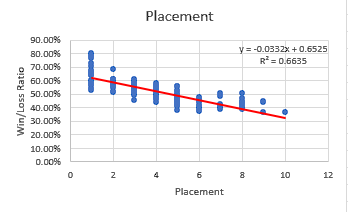
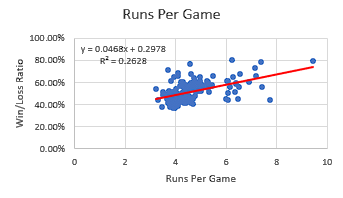
**Offensive Loss Scatter Plot 5 (graph 33)**



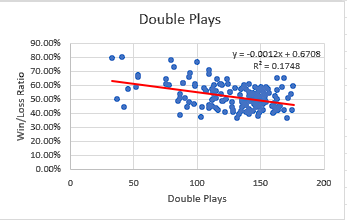
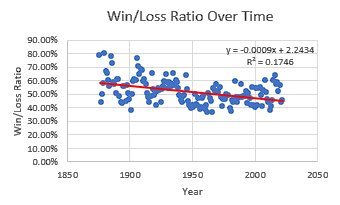
Similar to defensive stats, we see higher r-squared values in the loss correlations compared to the wins.

**Top 5 Offensive Win/Loss Ratio Correlation Graphs**

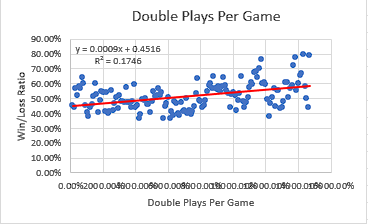
**Offensive Win/Loss Scatter Plot 1 (graph 34) Offensive Win/Loss Scatter Plot 2 (graph 35)**

** **

**Offensive Win/Loss Scatter Plot 3 (graph 36) Offensive Win/Loss Scatter Plot 4 (graph 37)**

**Offensive Win/Loss Scatter Plot 5 (graph 38)**

****

Similar again to the defensive stat patterns, we see placement has a significantly higher r-squared value compared to the other correlations, and we get some shockingly low r-squared scores, showing again overall poor fit / accuracy of correlations.

**Interpreting Results and Usage**

Looking at the scatter plots, we see a lot of poor correlations, the “loss” statistic in both defense and offense were the only graphs and correlations that would be deemed as accurate in terms of significance and accuracy. These scatter plots also prove the point mentioned above that despite statistical significance found in every variable tested, few of those variables had strong correlation and accuracy. We would traditionally like to see r-squared values above 0.8 and a strong accuracy would be above 0.9, which would allow us to say with certainty that there is a correlation strong enough to use reliably as a prediction.

With the results found so far, we would point the Cubs organization that defensive losses, and offensive losses is where our focus should be in terms of improvement. For our defensive data, it’s shown that runs, hits, and walks are the sole focus from our top findings. Our recommendation could be to focus on not allowing as many walks, especially if they are in a bases loaded situation, less walks would most likely lead to less runs which would increase our winning chances. With hits allowed we can recommend a different pitching sequence, developing a lead in counts to lead to chase pitches etc. For our offensive data, we see the more at bats, plate appearances, and double plays occurred, the more losses we have. Our recommendation would be to increase the number of hits, runs scored, effectiveness of our hitting attempts, with it being as poor as it is, that is why we see losses increasing with more attempts. Double plays would be hard to improve as it depends on how the ball is hit, we could try to work on stolen base success rate to not have a double play position as frequently.

We avoided some variables, for instance the more games played the more losses there are, this stat of games played shows up in a lot of places and has conflicting meanings at times. This variable is also something that cannot be fundamentally improved, as more postseason and world series chances, the more games played. We leave this in the findings but do not bother incorporating it into our discussion of improvements to the organization. It is also worth noting for example, the more plate appearances there were, was more correlated to offensive losses than wins. This was something I thought would correlate stronger the other way.

This gives us two possibilities; one is that predicting wins and losses with variables isn’t viable in the sports world. This is simply untrue, as even with low accuracy in correlations this doesn’t mean our results were pointless, this points us in the direction of a second possibility. This is that wins, and losses are built from almost all, if not all variables in the data, or with stats that aren’t calculated in the raw data and frequencies I had calculated. Luckily there is a linear regression method that will be used in python that will assist in determining variable calculation and prediction accuracy.

The method of analysis in python is to run a linear regression with the top 5 variables in correlations found, use those to predict how many wins/losses there will be with numbers we enter in. We can also calculate the r squared score of the regression analysis, this will tell us if it is more accurate or less accurate than the correlations and scatter plots found earlier. We will not be calculating win/loss ratio in this analysis as we saw in the scatter plots that variable had low r squared scores and even if it is marginally improved, wins and losses will be our best predictor.

With comparing 5 variables to one, this should increase our accuracy, as it was found it did. I ran tests with two, three, four, and five variables; this was done as I wanted to ensure increased accuracy the more variables added in. With this finding it proves the speculation that multiple variables have an outcome on wins and losses, we cannot predict this from one variable. With the regression ran in python, we can also build a predictor that allows us to input whatever number as each variable to see how many wins / losses we expect to see. Attached below are some outputs of what was ran with sample values inputted:

**Linear Regression Predictions (table 15)**

Text

Description automatically generated

We can see in this regression output our r squared values are consistently higher than what was found in the scatter plots, and within 0.1 of each other. Having a 75-85% accuracy here is not bad at all, if we were to add in more variables, which could easily be done, we would see the r squared values increase even further. I was surprised to see that defensive wins had a higher r squared score when the losses showed higher correlation in the Pearson values and r squared values on scatter plots. This is a more accurate representation of findings with multiple variables being compared at once.

There is one more step I wish to take in this project, and that is to split the data into training and test sets so we can see how accurate predictions can be in other data related to this structure. This can be done easily in Orange running linear regression with each variable. Since we are introducing more variables, we expect to see higher accuracy even with introducing a test set. All the data sets are divided into 66% training set, and 34% test set data. This is to allow most of the data to be focused on the dataset at hand, but allow new data to be introduced to strengthen our predictive accuracy if it’s found to be significant. The findings of each regression analysis are shown below:

**Defensive Win Predictions (table 19) Defensive Loss Predictions (table 20)**

** **

**Offensive Win Predictions (table 21) Offensive Loss Predictions (table 22)**

** **

Seeing that we were able to make a predictive model using all variables in a 98% accuracy for offensive and defensive wins and losses shows a pattern mentioned earlier; the more variables being compared, the higher accuracy we see in predictive models. This shows that the Cubs should focus on improving each stat we are looking at in this analysis. Even though some are more important than others, it’s important to have well rounded skills to allow for the best results. The predictive modeling could be useful if the organization wanted to see how the win/loss ratio would look with x amount of hits, runs etc. It’s great to see a predictive model be within 2 percent of a perfect score, this is a rarity to see in analysis, but reassures that baseball functions from other stats, dependent on each other for results.

This project provided a lot of information on the original questions asked and helped provide information that could be provided to the organization for specialized training and areas to focus on to improve their win/loss ratio and ideally get more postseason chances. The python regression analysis shows that with 5 variables we can predict wins / losses with a high accuracy. The only issue in this entire project is variables like “finish” and “games played” being in this analysis as this will increase our scores but doesn’t allow us to give feedback as these variables cannot be improved. If this were in a professional setting, we could suggest getting different data to still calculate from a lot of variables but get rid of the variables that have shown to not be useful in answering the original questions.

The final regression analysis shows that more variables lead to a near perfect analysis, which is as perfect as it gets in an analysis world; dealing with a 98-99 percent accuracy tends to be an anomaly. This shows that the data being kept is useful as everything coordinates wins and losses well. Seeing we were able to build a strong predictive model, we could also suggest to the organization we could find the team we have the worst win/loss ratio against and build a predictive model in attempts to see how to improve our wins against that team, or multiple teams. There are a lot of other usages of this concept of predicting wins and losses that can be done upon further investigation, but this provides solid groundwork for the questions wished to be answered at the start of this project.