Andrew – Connecting the Game

Analyzing the 30 MLB Franchises Since 1998

1. Introduction

1.1 Dataset Background

For my final project, I am analyzing a dataset containing information about all 30 Major League Baseball (MLB) franchises over the span of 1998-2021. I chose this specific time span because 1998 was the first MLB season when all 30 current active franchises were in the league and 2021 was the most recent full season.[1] To collect my data, I used two sources. The first and primary source was Baseball Stathead (Stathead) from Sports Reference.[2] This online statistical research tool allowed me to sort through each franchises' batting, pitching, and general statistics over the desired timespan I am working in. Stathead gave me all the variables in my dataset except for "MarketSize2012" which I found from Bleacher Report's article on teams' market size in 2012.[3]

1.2 Variables

Next, here is a list of the variables I am using for each team in my analysis and each variable's brief explanation. [4] I have included both my dataset and a more in-depth explanation of the numerical, technical baseball statistics I am using on my GitHub repository for this project.[5]

Table 1. List of variables and their basic descriptions

Lg	Which league a team is in (AL, NL, or Both)[6]
MarketSize2012	How big a team's market is (Small, Medium, or Big)

[1] https://www.baseball-reference.com/teams/#all_teams_active; See "Active Franchises" section.
[2] https://stathead.com/baseball/
[3] https://bleacherreport.com/articles/961412-mlb-power-rankings-all-30-mlb-teams-by-market-size
[4] I plan on excluding "Tm" from my analysis because both a team's name does not have any significance to a team's performance.
[5] https://github.com/Baseballfan5303/DataSciFinalProject, See "mlbTeamStats.csv" and "VariableExplanation.txt."
[6] AL stands for American League, NL stands for National League.

Salary	Compares how much money a team spent on its players compared to the league average
tmW-L%	A team's win-loss percentage
TB/G	Total bases per game
RC/G	Runs created per game
BAbip	Batting average on balls in play
BA	Batting average
OBP	On base percentage
SLG	Slugging percentage
OPS	On base percentage plus slugging percentage
OPS+	OPS as a normalized number across the whole league; considers factors like ballparks
hWAR	The total Wins Above Replacement (WAR) of every position player on a team
RA9	The number of runs against a team per nine innings
pWAR	The total WAR of every pitcher on a team
ERA+	Earned Run Average as a normalized number across the whole league; considers factors like ballparks

1.3 Visualization

Regarding data visualization, I generated a multitude of scatterplots which display the relationship between each of my numerical variables with one another. The full plot output can also be found on the same GitHub repository as mentioned before.[7] Additionally, here are some charts which will provide some visualizations of both important variables and groupings within my dataset.

[7] See "datasetPairsOutput.png."

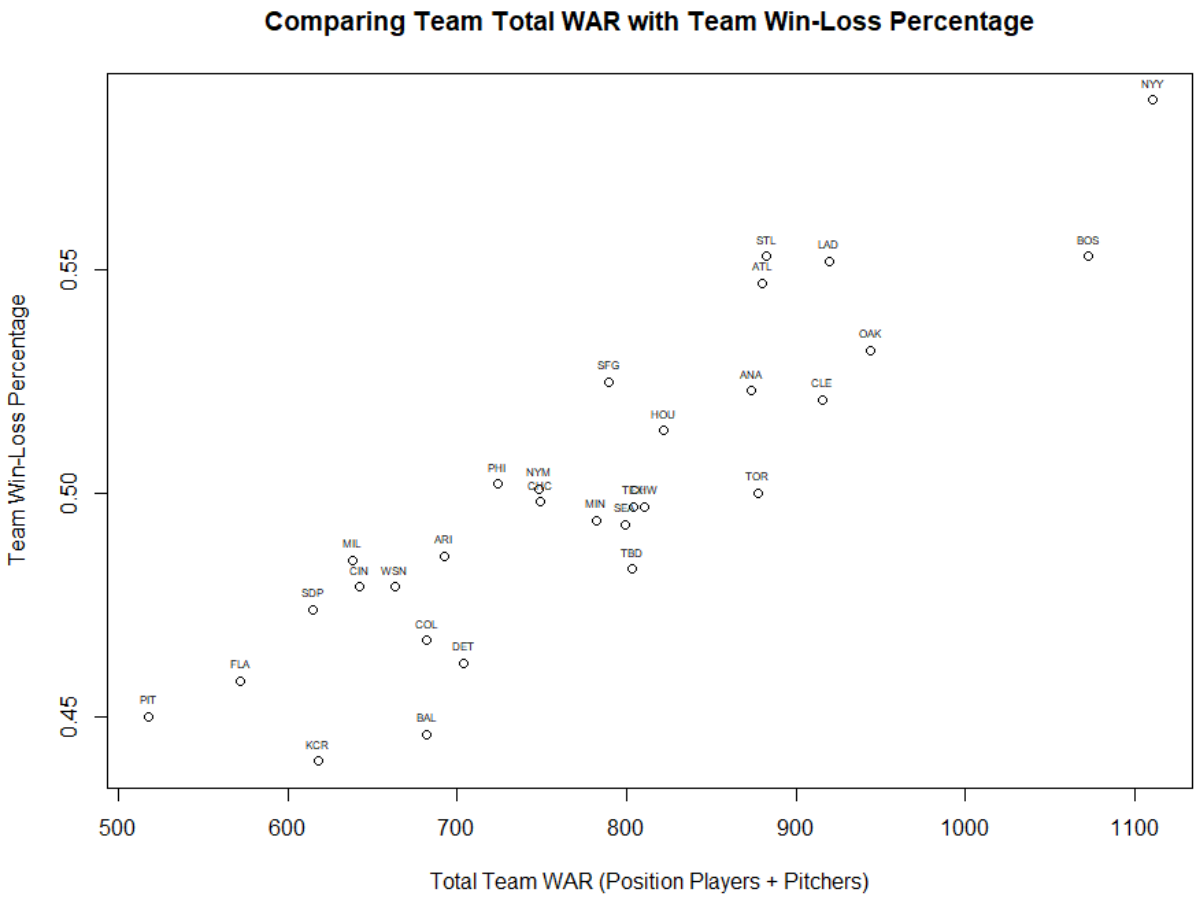

Figure 1. Comparison between a team's total WAR and their win-loss percentage.

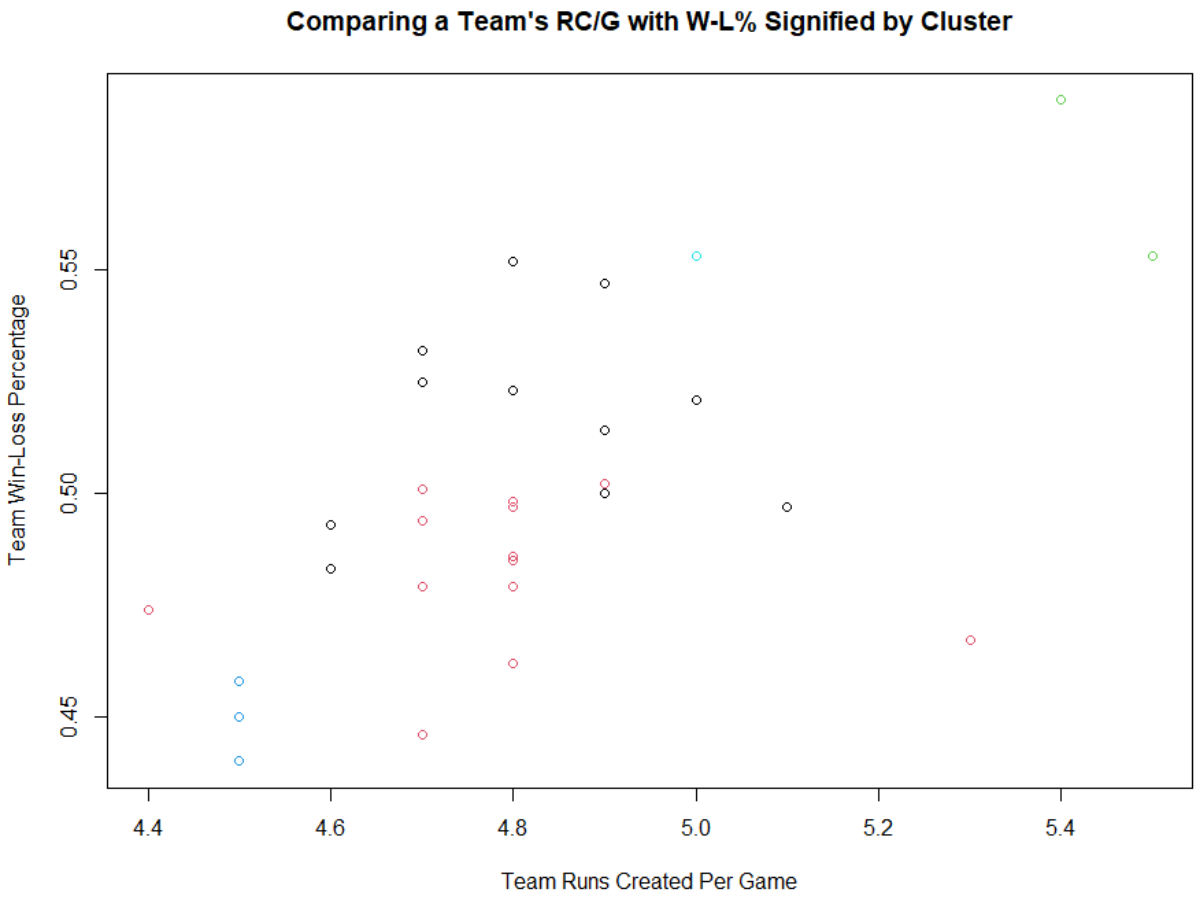

Figure 2. Comparison between a team's runs created per game and their win-loss percentage.

To discuss these scatterplots further, Figure 1 shows that there is a relatively strong, positive relationship between a team's total WAR and their success. Please note that the total WAR statistic was calculated by taking the sum of a team's hWAR and pWAR and was made only for the purpose of making this plot. When I create my models, I will be considering hWAR and pWAR separately. In addition, Figure 2 displays a slightly strong, positive relationship between a team's RC/G and their success. This plot also distinguishes sample units by what cluster they belong to, meaning which group a team belongs to based on similar features. Later in this report, I will reveal which teams are in each cluster and what qualities they share.

1.4 Goals

So, to analyze this dataset and make the results of my analysis relevant, I plan to utilize both unsupervised and supervised learning. First, I will use unsupervised learning to make a

dendrogram using hierarchical clustering. This will allow to me to determine which franchises are the most similar, to figure out if there any specific groups of franchises with similarity in their team information and statistics, and I hope to even ask the question of why this grouping might exist. Then, I will use regression to make a model which can predict a team's win-loss percentage based on rest of my dataset, which would include all the descriptive and statistical variables. A practical use for this regression model could be to predict how well a team will perform over a 23-year timespan depending on predictors such as, primarily, which league it ends up in, its market size, and even hypothetical baseball statistics given the team's performance. If we can relatively accurately predict a make-believe team's performance, this regression model could be used to determine if a new expansion team would be successful in the modern-day MLB.[8]

2. Methods

2.1 Clustering

In my opinion, clustering is the best technique for analyzing my dataset because placing the MLB franchises in groups is not only my main goal for this project, but also is a great indicator of which teams encountered similar success. To visualize clustering, I used the method of hierarchical clustering to produce a dendrogram, a tree-like plot which groups dataset units based on dissimilarity. Using hierarchical clustering and a complete linkage method will give us distinct groups of teams where each group is as much different than the next as possible.[9] Additionally, I placed boxes around each cluster to give a clearer visual distinction of the grouping occurring between the units. In my case, I have broken the franchises up into eight groups, which I found as the optimal number of groups using K-means clustering and the elbow point method.[10] The dendrogram I produced can be found in Section 3.1.

2.2 Linear Regression

---

[8] There has been some light discussion, on sites like Overtime Heroics and Fan Sided over the past two years, on which cities would be best suited for hosting MLB expansion teams.

[9] Complete linkage can be defined as "maximum between-group dissimilarity. Compute all pairwise distance between units in cluster A and units in cluster B and record the *largest* distance." (Liu)

[10] K-means clustering is another method for grouping units in a dataset. The elbow point method involves finding the *k* number of clusters where the total within-group distance begins to flatten out compared to other higher *k*-values.

Next, I developed a linear model for my MLB franchise dataset which could be used to predict the success of an expansion team.[11] This model is in the form of an equation, $y = a + bx$, where y is what we are trying to predict, a is the intercept, b are the statistics and information about a team, and x are the predictors or variables we use to predict. In this scenario, y is a response variable which will equal a team's predicted win-loss percentage based on all the b values of our predictors plus the value of a. Remember, the variables in this dataset are ones such as MarketSize2012, BA, OPS, etc.

2.3 Classification

Classification is identical to regression but instead forces the response variable to be a binary categorical variable. I plan on tackling this problem by setting my response variable to be determining the success of a team where 0 equals a team was unsuccessful, meaning their win-loss percentage was less than 50% and 1 equals a team was successful, meaning their win-loss percentage was greater than or equal to 50%. This process will act very similar to my goal for using regression, as both models will, in theory, predict the win-loss percentage for a given team. The only difference is that the regression model will output a specific percentage while the classification model will output a categorical result, whether a team was successful or unsuccessful.

3. Results and Analysis

3.1 Clusters

[11] Basically, I used a linear regression model generator where you input your response variable, predictors, and dataset to output a linear equation.

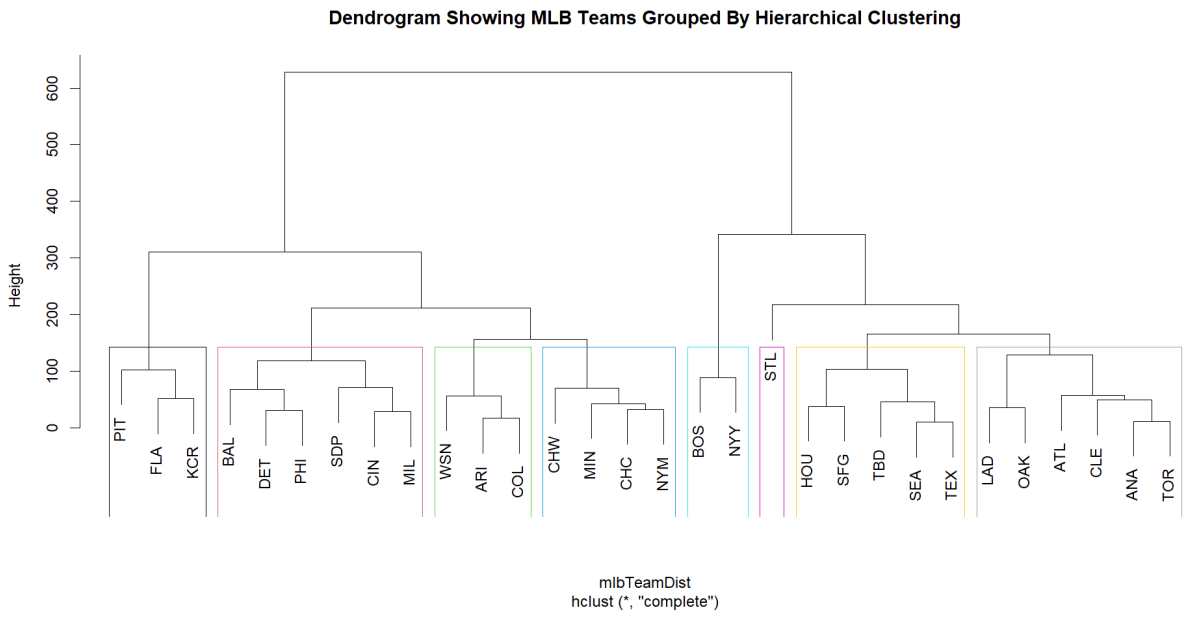

Figure 3. A dendrogram produced from my hierarchical clustering model.

Figure 3 groups the 30 MLB franchises from 1998-2021 based on maximum between-group dissimilarity. The teams are on the x-axis, broken up into eight distinct groups, and the height (y-axis) represents the distance between each group on the tree. Table 2 lists which teams are in each cluster and denotes their defining qualities.

Table 2. Lists each cluster and its defining qualities

Cluster Number	Teams in Cluster[12]	Features
1	PIT, FLA, KCR	-Smaller markets with below average salaries -Low team win-loss percentage -Low offensive production -Low wins above replacement
2	BAL, DET, PHI, SDP, CIN, MIL	-Varied markets and salaries -Low team win-loss percentage

[12] On my GitHub, I have included a text file explaining each team's initials used in both this table and in my visuals.

		-Lower offensive production -Higher offensive wins above replacement than Cluster 1
3	WSN, ARI, COL	-All National League teams with bigger markets -Below .500 team win-loss percentage -Higher pitching wins above replacement than both Cluster 1 and 2 -Above average pitching production
4	CHW, MIN, CHC, NYM	-Very close to .500 team win-loss percentage -Higher pitching wins above replacement than Clusters 1-3 -Above average pitching production
5	BOS, NYY	-Both American League teams with big markets -Very high team win-loss percentage -Highest offensive and pitching production -Highest offensive and pitching wins above replacement
6	STL	-Very high team win-loss percentage with a big market -Above average offensive production

		-Very high offensive wins above replacement -Low pitching wins above replacement
7	HOU, SFG, TBD, SEA, TEX	-Varied markets and salaries -Below average offensive production yet high offensive wins above replacement -Relatively good pitching production and pitching wins above replacement
8	LAD, OAK, ATL, CLE, ANA, TOR	-Varied markets and salaries -Around average offensive production yet very high offensive wins above replacement -Great pitching production and high pitching wins above replacement

3.2 Predicting

The output of my regression surprised me, to say the least. There are no significant variables, which shows that there is too high of a probability that for any of the predictors, they do not have a significant relationship with our response variable. However, because this dataset is technically considered a "small" dataset in the field of data science, there could be significant variables, but they just do not show up in the model since the pattern is not strong enough.[13] Yet, while there are no significant variables, the adjusted R-squared value is extremely high, and the p-value is extremely small. The R-squared value of 0.9495 tells us that about 95% of the

[13] 30 sample units is usually agreed upon as the number of sample units to be considered a "small" dataset. (Liu) The pattern in this case is the relationship between the response variable and the regression model.

variation in a team's win-loss percentage can be explained by our linear regression model. Moreover, the low p-value means that our data is both significantly significant and an indicator that there is a relationship between team win-loss percentage and the predictors incorporated.

In general, the classification model usually shared the same results as the linear regression model, just with a different interpretation which tells a better story. For example, if a team is in the National League or is present in both leagues, they are more likely to be unsuccessful. Teams with a smaller market tend to perform worse, but not necessarily in a way which will associate them with being unsuccessful; it might just mean their win-loss percentage is not too much higher than 50% compared with a medium or big market team. Furthermore, teams have a higher rate of success when they have higher RC/G, BA, and OBP. Teams are likely to be negatively affected when they have higher BAbip, RA9, and ERA+.

Meanwhile, the classification and regression models do have their differences in meaning when determining the impact of other variables not mentioned in the paragraph above. For instance, having a below average salary, higher TB/G, and higher OPS+ can associate with a lower team win-loss percentage, yet a team can still be successful. On the contrary, having higher OPS, hWAR, and pWAR relates to a team having a higher team win-loss percentage, but a team can still have unsuccessful results.

4. Conclusion

Finally, I would like to make some final remarks pertaining to this project going forward. First, I had to use some partial self-discretion when determining if a team had a small, medium, or big market. The most recent and available article I found ranking teams by market size was Bleacher Report's article. However, this was written in 2012, which might not necessarily reflect a team's market size right now. Yet, I hope that this market size will serve as the average for a team over the timespan I am analyzing, and I do believe that most teams' market size will only change if a team's performance drastically increases over a long period of time. I also had to trust Bleacher Report's rankings which do not include any monetary values, just discretionary rankings based on their analysis.[14] In addition, I initially believed that this dataset would only be

[14] In their article, they list nine teams as big market, eleven teams as medium market, and ten teams as small market. Even though their report is not visibly backed with data, I trust their reporting and had my dataset reflect their rankings.

ideal for clustering and that it would be impossible to use team statistics such as the ones I am using for predictions. See, how likely are we to realistically predict a team's exact, complicated baseball statistics over a span of 23 years? The idea of using this dataset to predict how well an expansion team will perform seems extremely difficult to accurately predict but also very intriguing. Hopefully as I increase my knowledge of data analytics and baseball statistics, I could hypothesize a way to formulate these enigmatic numbers.