Introduction and Dataset Description

The purpose of this project is to analyze the statistics of MLB Pitchers from the 2022 MLB season. The key attribute focused on in this project is **Adjusted Earned Run Average (ERA+)**. ERA+ is used by the MLB and normalizes a players ERA in relation to the rest of the league. The idea of this metric is to neutralize variables such as opponents faced or stadiums played in to compare pitchers.

The MLB's listed formula for this stat is: **Adjusted League ERA * 100 / ERA**. A perfectly average player will have ERA+ of 100, and a higher score is better. For context, the greatest ERA+ of all time (over the course of a career) is 205, held by Mariano Rivera. This value means that Rivera was 105% better than his average contemporaries [1].

This project's goal is to perform exploratory data analysis, compare ERA+ to other stats, and construct a supervised machine learning model for ERA+ prediction. The analysis aims to provide key insights that would interest MLB decision makers in evaluating pitcher performance.

The dataset analyzed in this project is the 2022 MLB Player Stats – Pitching [2], acquired via Kaggle. The original dataset contains 1081 rows and 35 columns, corresponding to players and their team or statistical information, such as Innings Pitched (IP), Hits Allowed (H), and Walks (BB). The dataset has no missing or N/A values.

The following adjustments were made to the data, along with the justification for each:

- Adjusted Name attribute to remove unusual characters and ensure readability
- Manually added a "Throws" attribute to track handedness. This will allow for distinction between the two types, and was not contained in the original data
- Players who were on multiple teams in 2022 had multiple entries. Only their season long stats were selected, and the Team/Div/League attributes were adjusted to reflect if they played on multiple teams.
- Dropped Earned Runs (ER) and Earned Runs Average (ERA) attributes. These variables have high correlation with our eventual target variable, and the essence of their information is still contained in ERA+. Given that our main focus is ERA+, we do not need these predictors.
- Removed data with ERA+ equal to zero, as this is not meaningful. A pitchers standard ERA would have to be infinite (based on the MLB's formula) to get ERA+ of zero, and that would only happen in rare cases that are not relevant to this study. This included 53 observations.

Exploratory Data Analysis

ERA+ Distribution and Outlier Discussion

As stated, the highest ERA+ over a career is 205. From this fact, it is reasonable to assume that any ERA+ over this value could be regarded as a statistical outlier. In this dataset, there are 36 players that present ERA+ > 205, denoted by the players to the right of the red line in Figure 1, and are prime candidates for statistical regression. While not exhaustive, the following list explains the reasoning for some players having outlier ERA+ measurements:

1. Elite-level reliever. The top closers in baseball are relied upon for "Save (SV) situations", generally defined as the last inning of the game where the pitchers team is winning by 3 runs or less, allowing less margin for error. The best of the best closers or high-leverage relievers may exceed this limit in a select season. This is the category Mariano Rivera falls into.
2. Extremely small sample size. Pitchers who have very few innings or appearances may have great success, but given more volume should return to a more normal value.
3. Abnormally "lucky" season. This is less likely given a full seasons worth of stats, however, it is possible that a pitcher had more good fortune than normal, without experiencing regression. More often than not, this is not sustained over a career.

Outliers aside, the remaining data is relatively normally distributed about 100. This is expected given the definition of ERA+, where 100 refers to what should be average. There is slightly more volume on the lower side, however, this is a result of offsetting the outliers in the 205+ range.

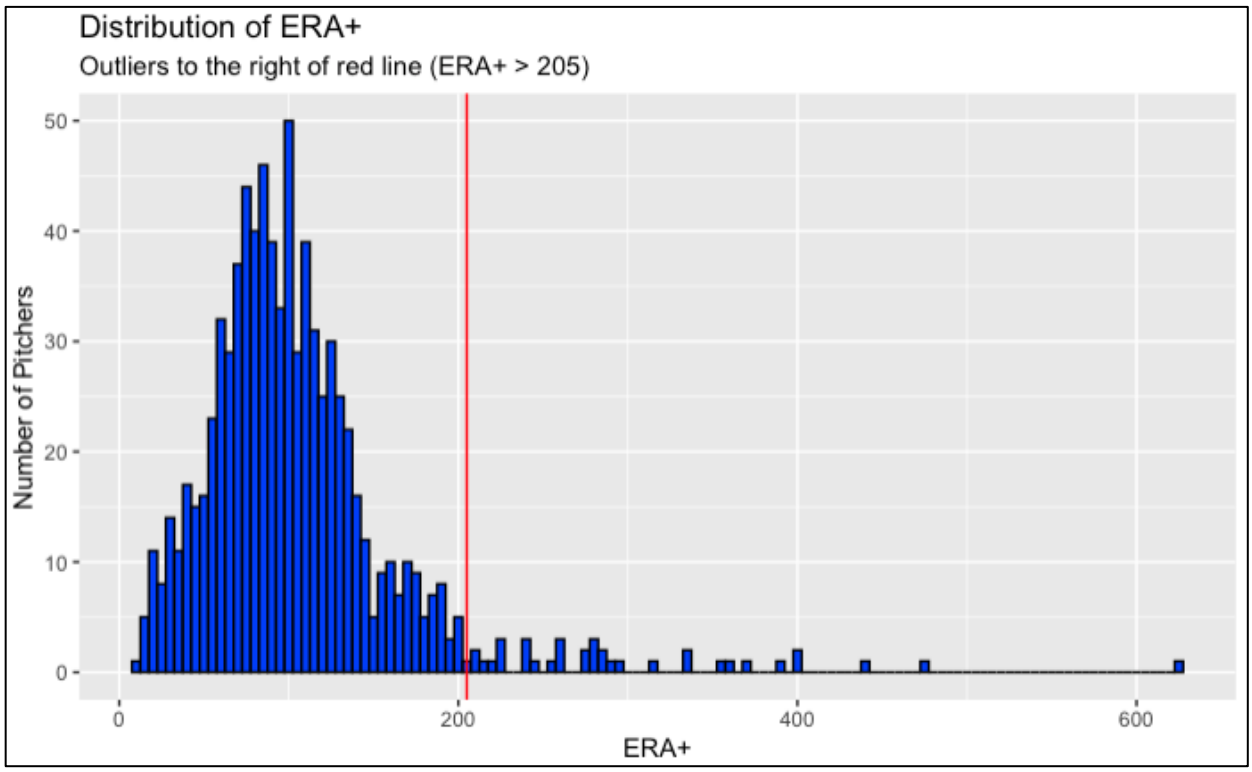

Figure 1

Principal Component Visualizations

In Figures 2 and 3, the observations are plotted versus the first two principal components. In Figure 2, the data is colored by ERA+ value, while in Figure 3, the colors refer to custom kmeans cluster centers, defined by:

1. Red - Starting Pitchers (SP). Most Games Started (GS)
2. Green - Relief Pitchers (RP) and Closing Pitchers (CP). Most Games appeared in (G)
3. Blue - Other / No clear distinction.

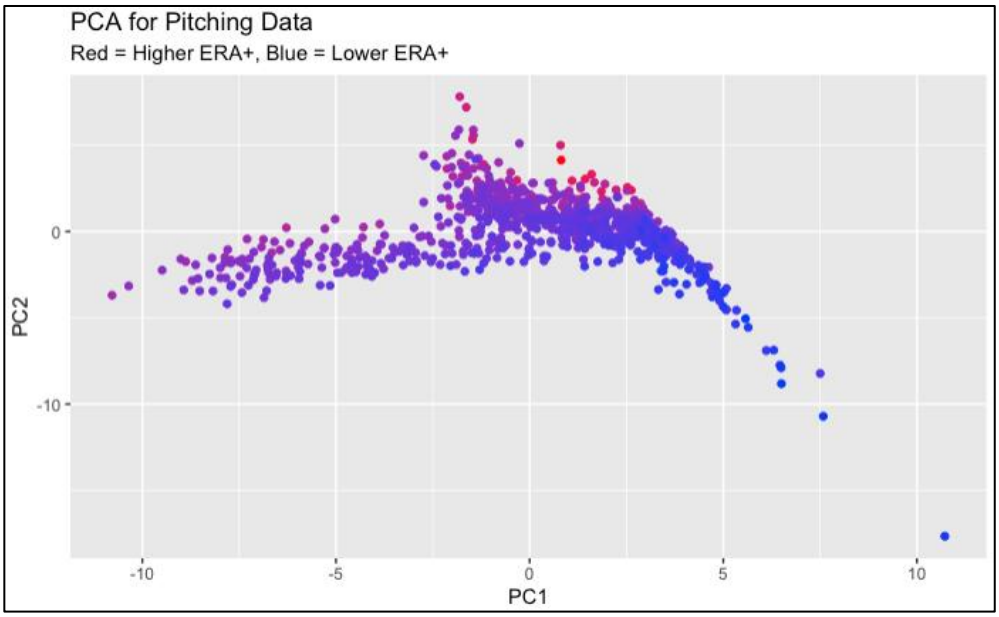

Figure 2

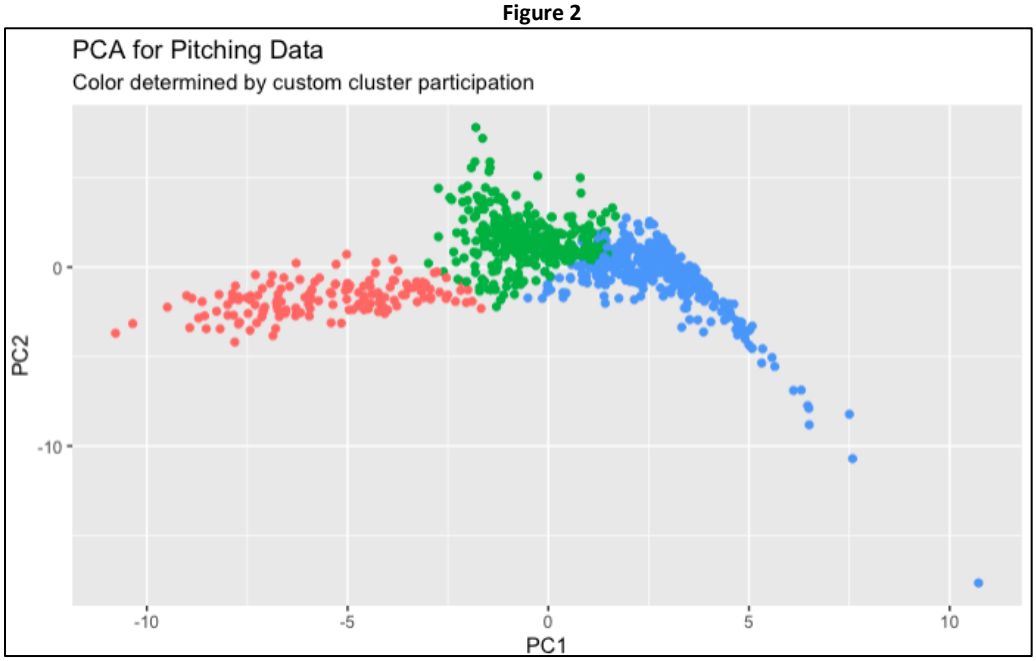

Figure 3

From Figure 2, it is clear that the players with better ERA+ values have higher PC2 values, and middling PC1 value, which corresponds to the relief pitcher cluster (green) of Figure 3. The cluster of starting pitchers (red) has mostly purple fill in Figure 2, indicating a middling ERA+. This matches what I would expect from domain knowledge. Starting pitchers are relied upon to pitch a handful of innings each appearance, exposing themselves to more opportunities to accumulate Earned Runs. While there are elite SP who may keep their ERA+ low (light purple fill in Figure 2), this is not a shared trait of the entire cluster. The remaining cluster (blue in Figure 3) is where most of the low ERA+ values can be observed, although there are a handful that have higher ERA+, particularly closer to the cluster 2/3 boundary line. This cluster is where many of the low-end pitchers are, including position players, minor league fill-ins, or players with extremely limited appearances.

Variable Correlation

For numeric predictors, the full correlation matrix can be observed in Figure 4. The key observations that are important for this analysis are:

1. WHIP (Walks + Hits per Inning Pitched) is highly correlated with H9 (Hits per 9 innings pitched) and BB9 (Walks per 9 innings pitched).
2. The three variables most correlated to ERA+ are WHIP, H9, and FIP (Fielding-Independent Pitching).
3. ERA+ has very little correlation with Wins and Losses.
4. Statistics that accumulate with more pitching opportunity (Innings Pitched (IP), Strikeouts (SO), Runs (R), Batters Faced (BF), etc.) have high correlation. The raw counting stats are more of a reflection of volume, rather than success.

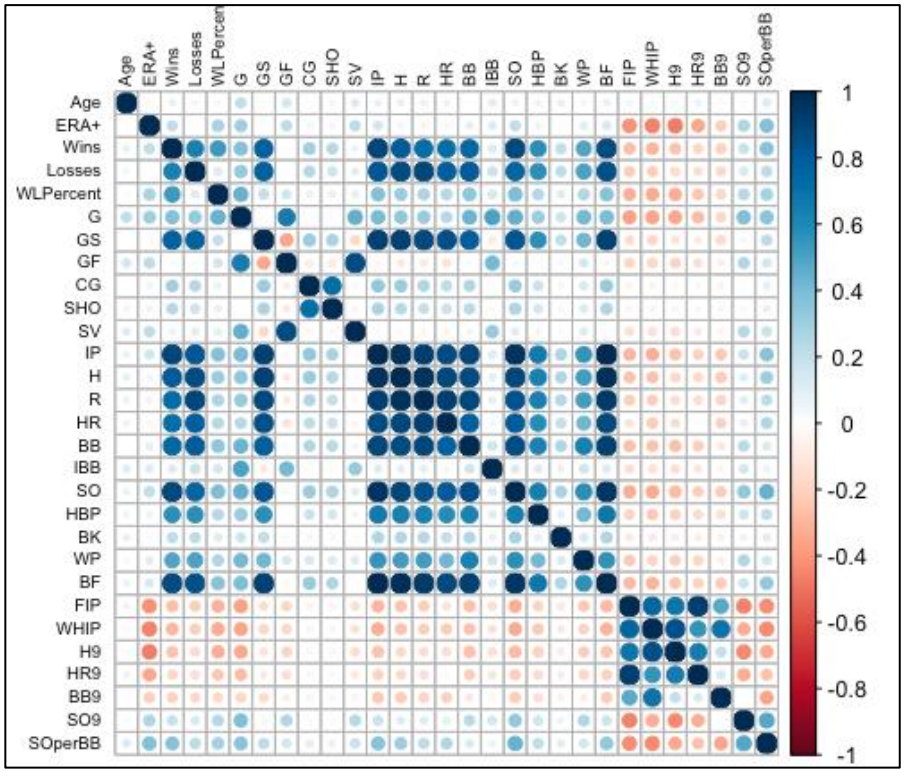

Figure 4

ERA+ against Select Predictors

In Figure 5, we can observe the how ERA+ is distributed by MLB Team. Unsurprisingly, perennial contenders (and high spenders) such as the Dodgers (LAD) and Yankees (NYY) find themselves at the top of the league, along with the Astros (HOU), who were the eventual World Series champions. On the other side of the spectrum, teams who finished near the bottom of the league in 2022 can be seen on the low end. The Athletics (OAK), Pirates (PIT), Nationals (WSN), and Royals (KCR) are the four worst teams in ERA+, and coincidentally had four of the worst records in the MLB regular season. Overall, out of the 12 teams that made the postseason in 2022, only two teams had below average (< 100) team ERA+, being the Cardinals (STL) and Padres (SDP). This finding reinforces the importance of maintaining at least a league-average pitching staff if a team plans on competing for a championship, rather than relying on pure offense. If this is a trend that historically holds, ERA+ may serve as a reliable mid-season indicator of a team's postseason potential from an analyst or even gambler perspective.

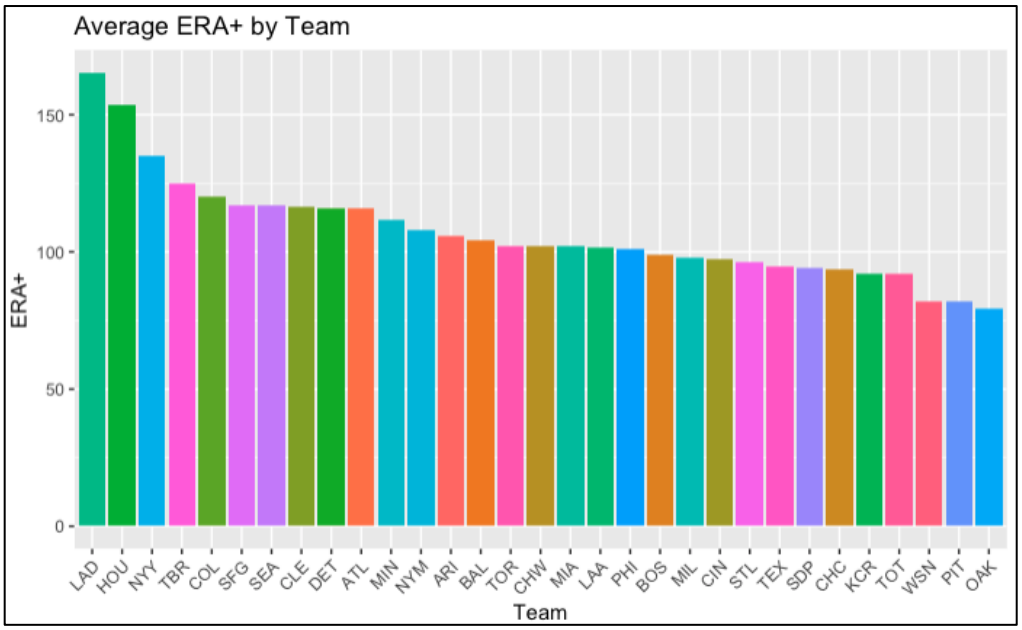

Figure 5

The relationship between ERA+ and Age is examined in Figure 6. Between ages 21-37, there is no clear trend indicating what age(s) may be a pitchers best. There is a wide belief that a player has a "prime", where youth and experience overlap resulting in their best years. The 2022 data, however, does not clearly reflect that. An uptick can be seen at ages 31-33, but it does not appear significant, and the highest average belongs to 37-year-olds. The one concrete insight gathered from this is that every pitcher aged 38 or older recorded a below average ERA+, indicating a potential trend for when to move on from an aging asset.

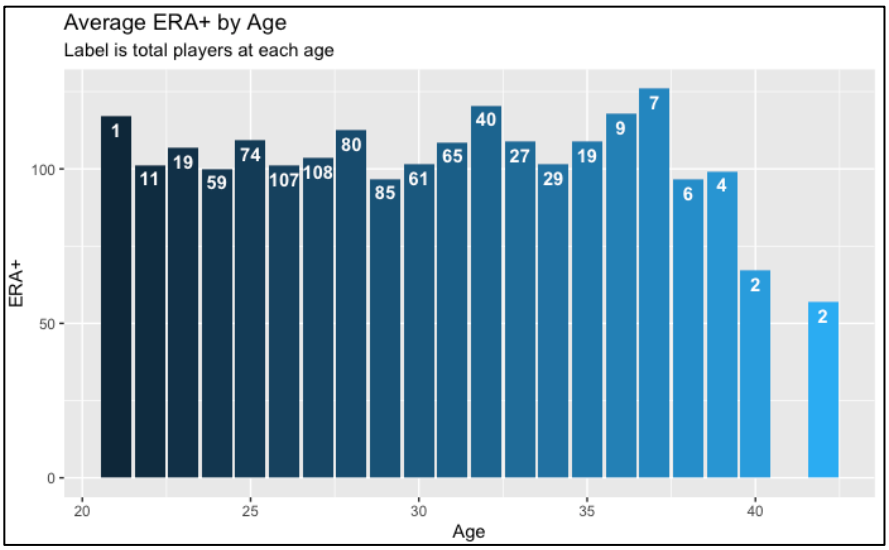

Figure 6

The last metric we will look at is ERA+ vs Handedness. The data in Figure 7 shows no clear advantage to throwing left handed (LHP) or right handed (RHP), as the ERA+ values are nearly level. Handedness, at least in the 2022 season, did not appear to play a significant factor in pitcher effectiveness.

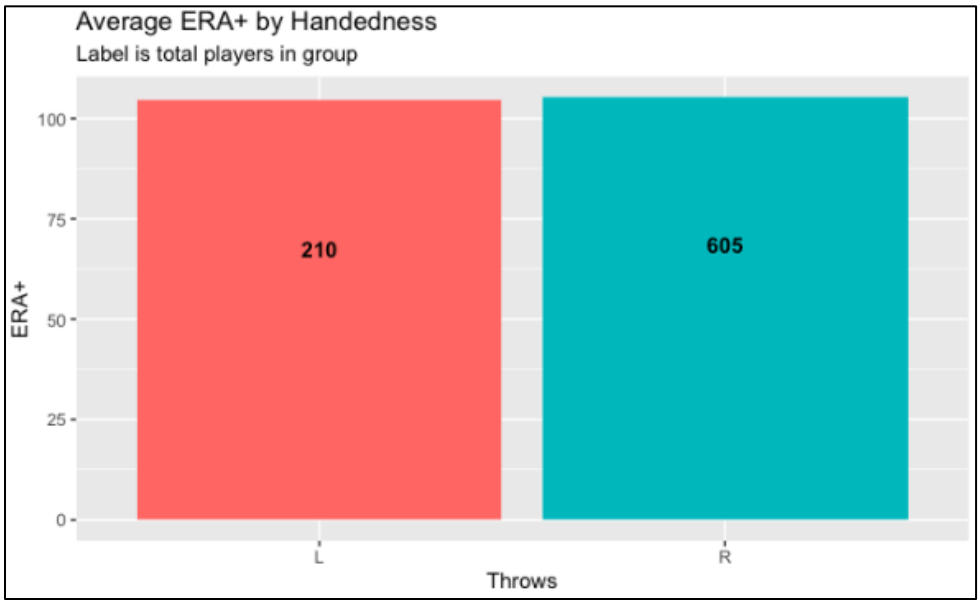

Figure 7

ERA+ Prediction

Feature Selection

There are over 30 distinct attributes in this dataset. Many are derivations of one another, resulting in some highly correlated predictors. Feature selection will utilize stepwise selection based on BIC to ensure this level of complexity is penalized. From this method, the 10 suggested final predictors are:

- Wins (W)
- Losses (L)
- Win/Loss% (WLPercent)
- Runs (R)
- Walks (BB)
- Strikeouts (SO)
- Batters Faced (BF)
- Hits per 9 Innings Pitched (H9)
- Walks per 9 Innings (BB9)
- Strikeouts per Walk (SOperBB)

Considering W and L are suggested to be tracked, I made the decision to remove WLPercent, as it is a direct redundancy. There is a case to be made that W and L are more team based and should be removed altogether, however, I think there is enough potential information to be learned from their inclusion. A stat of note that was omitted was WHIP. Given the relevance of WHIP to a pitcher's performance (measuring how many base-runners a pitcher allows per inning), I see justification to include it. Its inclusion replaces the need for both BB9 and H9, as WHIP encapsulates their information already and removes an extra dimension, while also giving a way to measure Innings Pitched. I also believe that FIP is likely critical when evaluating a pitcher, and decide to include it against the suggestion of the stepwise BIC selection process. Finally, three different versions of walks-based stats are suggested, being BB, BB9, and SOperBB. The inclusion of only one should be sufficient, and given BB9 was already removed above, I elect to remove SOperBB and keep only BB. Since SO and BB are already kept, SOperBB is redundant. These choices result in the final eight predictors below, with a corresponding dendrogram in Figure 8.

- Wins
- Losses
- Runs
- Walks
- Strikeouts
- Batters Faced
- Walks + Hits per Inning Pitched
- Fielding-Independent Pitching

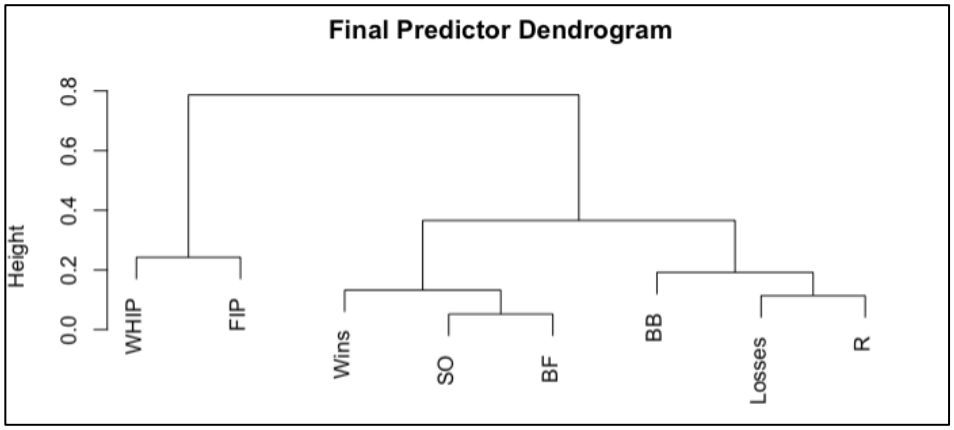

Figure 8

Data Preparation for Model Use

Before constructing and fitting any models, the predictors (all numeric) were scaled. This excludes the target variable of ERA+, given that its format is already a scaled and centered version of ERA, the center being 100. Preserving the context of the metric is essential to accurately compare predictions against actual values. Additionally, the data was split into 80% train data, 20% test data. Each model is trained and tested on the same splits of data, allowing a fair evaluation.

Models Used

Table 1 summarizes the models selected, as well as the justification for each.

Table 1

Model	Justification
Linear Regression (LR)	Provides a simple, interpretable baseline
Random Forest (RF)	Quick, great for non-linear relationships, allows tracking of variable importance
Support Vector Regression (SVR)	Uses kernel function to capture complex, non-linear patterns (albeit less interpretable)
Extreme Gradient Boosting (XGBoost)	Boosting algorithm that should improve upon Random Forest. Also allows feature importance tracking.

Model Results

Table 2 summarizes the model variations construction, as well as the RMSE, MAE, and R^2 for each.

Table 2

	Model	ModelNotes	RMSE	MAE	R.2
1	Linear Regression 1	Baseline MLR	54.57706	27.92153	0.3860267
2	Linear Regression 2	Used 5-Fold CV	45.12710	26.74070	0.4562429
3	Random Forest 1	ntree = 600, mtry = 5	43.44129	18.85381	0.6249807
4	Random Forest 2	ntree = 1000, mtry = 7	43.50486	19.02332	0.6202718
5	Random Forest 3	Bagging	43.33645	18.73468	0.6231234
6	SVR	radial kernel	41.08773	16.30716	0.6717771
7	XGB	nrounds = 100	39.79975	18.44049	0.6697623

Models 6 and 7 have similar performance, and their own arguments for best model. Since 7 (XGBoost) has minimal hyper-parameter tuning, lower RMSE (overall accuracy in prediction), comparable R2 (0.002 less than M6), and ability to view variable importance, it is selected as the top-performing model for predicting ERA+. Figure 9 shows how XGBoost's predictions compare to actual ERA+ values.

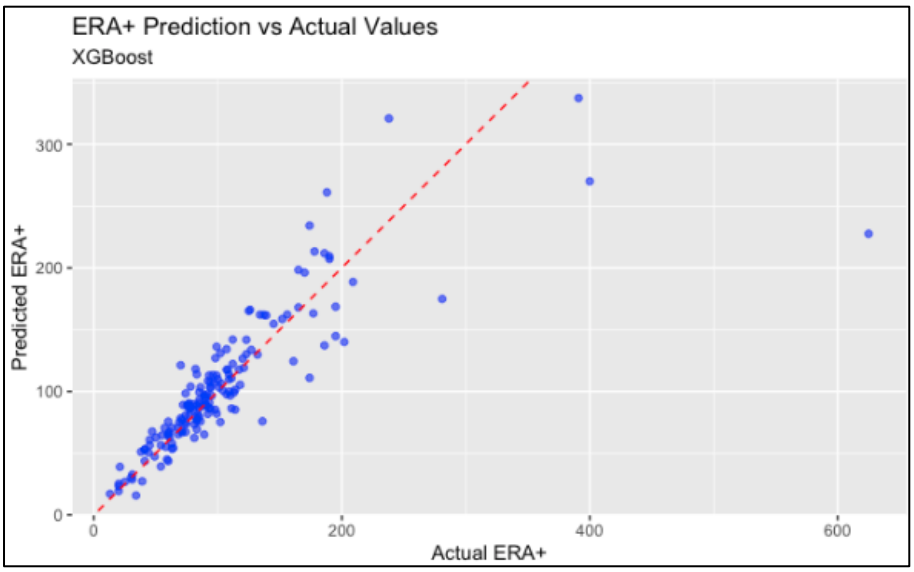

Figure 9

The model shows promise with prediction of ERA+ values between 0 and 150, while showing struggles at higher values, particularly above the outlier threshold. Since we have already established an ERA+ above 205 is possible but abnormal, it may be important to note how the model performs for "normal" values versus those outliers. The mean prediction error for outliers was ~131, while it was only ~14 for the other values. This fact leads me to believe the model has more utility than the overall metrics may indicate, as the current model shows ability to predict ERA+ within ±15 points for typical observations.

In Figure 10, the ordered importance of the predictors (as defined by XGBoost) can be seen. WHIP emerged as, by far, the strongest predictor. This reaffirms the intuition used to include it. FIP, the other manually selected attribute, finds itself as the 3rd most important, validating its relevance as well. Requiring less intuition than others, Runs (R) is also a highly valued metric. In the 2022 data, 91% of Runs are Earned ("Unearned" Runs are not attributed to the pitcher and do not factor into ERA+), making R an understandably huge factor in ERA+. Meanwhile, as noted during variable correlation analysis, Wins and Losses may say more about a team than an individual pitcher, and that is reflected in their variable importance being by far the lowest.

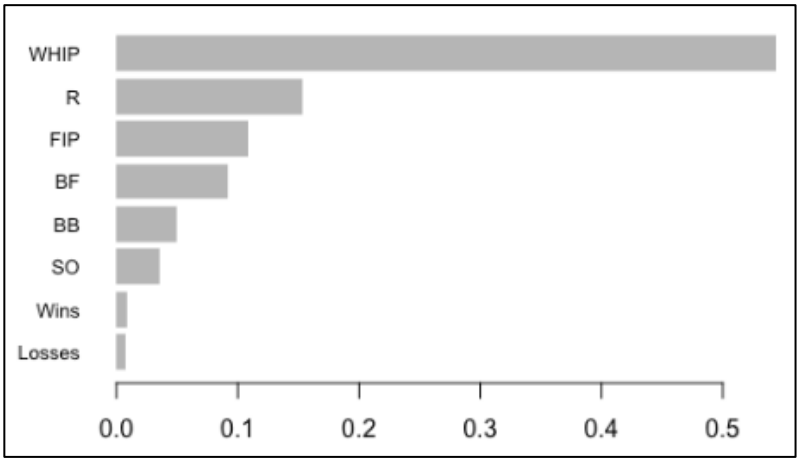

Figure 10

Conclusion

The analysis of this data explored the relationship between a pitchers ERA+ and other various statistics from the 2022 MLB Season.

Categorical predictors (Team, Age, and Handedness), revealed underlying information when compared to a pitchers ERA+. Clustering and principal component analysis revealed that ERA+ values may be used to distinguish a starting pitcher, relief pitcher, or non-MLB level pitcher. Observations made during correlation analysis held true throughout and supports final results.

Through feature selection, combining BIC analysis and domain knowledge, a group of eight predictors was selected for use in supervised machine learning models. Non-linear models proved more successful, with XGBoost being assigned the champion model. The models used allowed for analysis of predictor importance, with WHIP standing out the most, and Wins/Losses claiming near irrelevance.

The inclusion of data from past years would be a next logical step to improve this projects conclusions and machine learning models. Such data may allow the insights to be confirmed or adjusted given the history of the MLB.

Key Insights

- Pitchers with higher ERA+ profile better as RP or CP, rather than SP. Their success is better for high-leverage situations rather than a multi-inning outing.
- Team success can be influenced by the average ERA+ of their pitching staff. The four worst teams in the MLB had the lowest average ERA+, and only 2/12 playoff teams were below average in ERA+. Clubs should prioritize a league average staff, at minimum, to increase chances of success.
- Age and handedness do not have significant impact on ERA+ in isolation. Pitchers age 38+ displayed strong regression, but no clear trends emerged before then. Handedness has more utility in specific gameday matchups than season long prediction.
- Supervised learning models show moderate success in ERA+ prediction. XGBoost performs the best on this dataset, offering reliable predictions for the normal range of values (0 - 205), but experiences difficulty with outlier values (205+).
- Among predictors selected, WHIP stood out as the most critical. Team executives should refer to WHIP, above other metrics, to drive player analysis and influence high-stakes decisions. Wins and Losses offer little utility, and should not be a main consideration when evaluating a pitcher.

References

[1] MLB. (n.d.). Earned Run Average Plus (ERA+). Major League Baseball. https://www.mlb.com/glossary/advanced-stats/earned-run-average-plus

[2] Vinco, V. (2022). 2022 MLB Player Stats – Pitching [Data set]. Kaggle. https://www.kaggle.com/datasets/vivovinco/2022-mlb-player-stats?select=2022+MLB+Player+Stats+-+Pitching.csv

Appendix (Code Snippets)

Figure 1:

```r
{r}
ggplot(df) +
  aes(x = `ERA+`) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  geom_vline(xintercept = outlier_cutoff, color = "red") +
  ggtitle("Distribution of ERA+","Outliers to the right of red line (ERA+ > 205)") +
  labs(x = "ERA+",y = "Number of Pitchers")
```

Figure 2:

```r
prcomp_p <- data.frame(
  prcomp(x = df_scaled_numeric)$x[,1:2],
  Name = player_names,
  Rank = df$`ERA+`
)
ggplot(prcomp_p) +
  aes(x = PC1,y = PC2,color = Rank,label = Name) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red",) +
  ggtitle("PCA for Pitching Data","Red = Higher ERA+, Blue = Lower ERA+") +
  theme(legend.position = "none")
```

Figure 3:

```r
zeroes <- rep(0, length.out = length(colnames(df_scaled_numeric)))
starting_pitchers <- zeroes
starting_pitchers[which(colnames(df_scaled_numeric)=="GS")] <- 3
closing_pitchers <- zeroes
closing_pitchers[which(colnames(df_scaled_numeric)=="G")] <- 3
other_pitchers <- zeroes

initial_centers <- matrix(
  data = c(starting_pitchers, closing_pitchers, other_pitchers),
  nrow = 3,
  byrow = TRUE
)
rownames(initial_centers) <- c("starters","closers/relievers","other")


kmeans_customCenters <- kmeans(x = df_scaled_numeric, centers = initial_centers)
#kmeans_customCenters$centers

prcomp_custom <- data.frame(
  prcomp(
  x = df_scaled_numeric)$x[,1:2],
  Name = rownames(df_scaled_numeric),
  Rank = df_scaled_numeric$`ERA+`,
  Cluster = as.character(kmeans_customCenters$cluster)
)
ggplot(prcomp_custom) +
  aes(x = PC1,y = PC2,color = Cluster) +
  geom_point() +
  ggtitle("PCA for Pitching Data","Color determined by custom cluster participation") +
  theme(legend.position = "none")
```

Figure 4:

```
corr_matrix <- cor(df_numeric)
corrplot(corr_matrix,
     tl.col = "black",
     tl.cex = 0.6)
```

Figure 5:

```
df %>%
  group_by(Team) %>%
  summarise(Avg_ERAplus = mean(`ERA+`)) %>%
  ggplot(aes(x = reorder(Team, -Avg_ERAplus), y = Avg_ERAplus, fill = Team)) +
  ggtitle("Average ERA+ by Team") +
  geom_col(show.legend = FALSE) +
  labs(x = "Team", y = "ERA+") +
  theme(axis.text.x = element_text(angle = 45)
  )
```

Figure 6:

```
df %>%
  group_by(Age) %>%
  summarise(Avg_ERAplus = mean(`ERA+`), Count = n()) %>%
  ggplot(aes(x = Age, y = Avg_ERAplus, fill = Age)) +
  ggtitle("Average ERA+ by Age", "Label is total players at each age") +
  geom_col(show.legend = FALSE) +
    geom_text(
    aes(label = Count),
    vjust = 1.5,
    color = "white",
    fontface = "bold"
  ) +
  labs(x = "Age", y = "ERA+")
```

Figure 7:

```
df %>%
  group_by(Throws) %>%
  summarise(Avg_ERAplus = mean(`ERA+`), Count = n()) %>%
  ggplot(aes(x = Throws, y = Avg_ERAplus, fill = Throws)) +
  geom_col(show.legend = FALSE) +
  ggtitle("Average ERA+ by Handedness", "Label is total players in group") +
  geom_text(aes(label = Count),
    vjust = 10,
    color = "black",
    fontface = "bold"
  ) +
  labs(x = "Throws", y = "ERA+")
```

Feature Selection:

```
baseline_lm <- lm(`ERA+` ~ ., data = df_var_sel)
stepwise_model <- step(baseline_lm, direction = "both",
                       k = log(nrow(df_var_sel)), trace = FALSE)
summary(stepwise_model)
```

Figure 8:

```
final_preds_list <- c("ERA+", "Wins", "Losses", "R", "WHIP", "SO", "BF", "BB", "FIP")
df_final <- df_var_sel[final_preds_list]
final_preds <- df_final
final_preds$`ERA+` <- NULL
final_dist_matrix <- as.dist(1-abs(cor(final_preds)))
final_dend <- hclust(final_dist_matrix)
plot(final_dend, main = "Final Predictor Dendrogram")
```

Figure 9 and XGBoost implementation:

```
X_train <- as.matrix(train_data %>% select(-`ERA+`))
y_train <- as.matrix(train_data$`ERA+`)
X_test <- as.matrix(test_data %>% select(-`ERA+`))
y_test <- as.matrix(test_data$`ERA+`)

xgb_model <- xgboost(data = X_train, label = y_train,
  nrounds = 100,
  eval_metric = "rmse",
  verbose = 0
)

y_pred <- predict(xgb_model, X_test)

mae_xgb <- mean(abs(y_pred - y_test))
rmse_xgb <- sqrt(mean((y_pred - y_test)^2))
r2_xgb <- 1 - sum((y_test - y_pred)^2) / sum((y_test - mean(y_test))^2)

importance <- xgb.importance(model = xgb_model)

results_xgb <- data.frame(
  Actual_ERA_Plus = actuals,
  Predicted_ERA_Plus = round(y_pred,2)
)

xgb_plot <- ggplot(results_xgb, aes(x = Actual_ERA_Plus, y = Predicted_ERA_Plus)) +
  geom_point(alpha = 0.6, color = "blue") +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  ggtitle("ERA+ Prediction vs Actual Values", "XGBoost") +
  labs(x = "Actual ERA+",y = "Predicted ERA+")
xgb_plot
```