MLB Pitch Effort and Injury Analysis

Matthew Rice

2025-03-02

Introduction

Sports injuries are a concern for athletes and coaches at all levels. In the game of baseball, elbow injuries are a particular concern for pitchers, and at the highest levels team owners and managers have to factor in these concerns when making transactions and offering contracts. Data analysis in Major League Baseball (MLB) has driven the trend of pitchers attempting to increase both the velocity and spin rate of their pitches in order to obtain better outcomes. An important question being addressed in this project is: "How much (if any) effect does attempt to maximize velocity and spin rate have on MLB players' time spent on the Injured List (IL)?"

Research Questions

To better understand the relationship between pitcher effort and injury I'll begin with these questions:

1. What data is available on the pitch velocity and spin rate of MLB players?
2. What data is available on the frequency and length of trips to the IL for MLB pitchers?
3. Are there any variables used that can used to determine "effort" other than velocity and spin rate?
4. Should all pitches be included in the data set or only certain types?
5. Do any trends or patterns stand out when viewing summary statistics of the available data?
6. Can any variables be combined to create new variables that might be more concise in how they relate to injury time?

Approach

I first want to define "effort" before running tests that will give insight into how that variable is related to time missed due to injury. All athletes who reach the highest level of their chosen sport do so through an extreme amount of effort in how they train, study, and compete. For the purpose of this project is, I want to look specifically at how often MLB pitchers throw the ball as hard as they can. To ascertain this, I will be looking at maximum values of both pitch velocity and spin rate of pitches from a large selection of players over the past five years. I am going to examine how those max values compare to the means of pitch velocity and spin rate. And in that attempt to better define "effort", I will create new variables that measure the percentage of pitches thrown at or above 95% of the maximum value of velocity and spin rate. My hypothesis is that these variables will have an effect on time missed due to injury.

Data

The first table from baseballsavant.mlb.com provides average pitch velocity and spin rate for all MLB pitchers that have thrown at least 2,000 total pitches over the last five years. Opening each player profile individually accesses a .csv file with detailed information of each pitch thrown.

Next, data from fangraphs.com contains Injury List details for every MLB team by year. For my purposes data from each of the past five years was downloaded as an Excel file and combined.

The final table, also from baseballsavant.mlb.com, is similar to the first, but contains data from minor league baseball players, and will be used for further research.

```
data <- "/Users/macuser/Documents/final_pitch_data_correct.csv"

final_pitch_data <- read.table(file = data, header = TRUE, sep =",")
```

The data obtained from baseballsavant.mlb.com was a table with a size of 759,746 x 113. Each observation documented a fastball thrown by a Major League Baseball pitcher. The data was for total pitches thrown over the last five seasons (2020 - 24). The list includes pitchers who have thrown at least 2,000 total pitches over that time span, which came to a total of 196 players. To make the data more manageable I grouped by player name and selected three quantitative variables that were most pertinent to the problem being addressed. Those variables are: max_release_speed, average_velo, and average_spin_rate. I then created a new quantitative variable, "percentage_above_95". This value measures the percentage of fastballs thrown with a measured velocity between 95 - 100 percent of the player's maximum pitch velocity.

The injury data obtained from fangraphs.com included two variables titled "IL Retro Date" and "Return Date". Subtracting the retro date from the return rate created a new variable, "Days_missed", which I use as the dependent variable in my analysis. By using the left_join function, I combined these two data sets into a single table with a size of 196 x 6.

variable	class	first values
player_name	Character	Abbott, Andrew
max_release_speed	Integer	95.8
percentage_above_95	Integer	97.96
Days_missed	Integer	41
average_velo	Integer	92.76
average_spin_rate	Integer	NA

Required Packages

• rmarkdown: RMarkdown is an open-source tool that helps create dynamic reports or documents that enables to keep all codes, results, plots, and writing in one place.

• knitr: knitr packages provide user-controlled output for dynamic report generation from RMarkdown files.

• ggplot2: This is the most widely used versatile package for data visualization that provides a coherent system for describing and building aesthetic visualization using graphs.

• dplyr and plyr: This package is commonly used for existing data transformation into a better suited format for data analysis and visualization.

• Matrix: The Matrix package provides a powerful framework for creating, manipulating, and performing operations on matrices efficiently.

Plots and Tables Required

• Histograms: This is to show the distribution of values for a variable.

• Scatter plots: This is to show two variables in comparison with each other

• Density plots: This is to show the density distribution

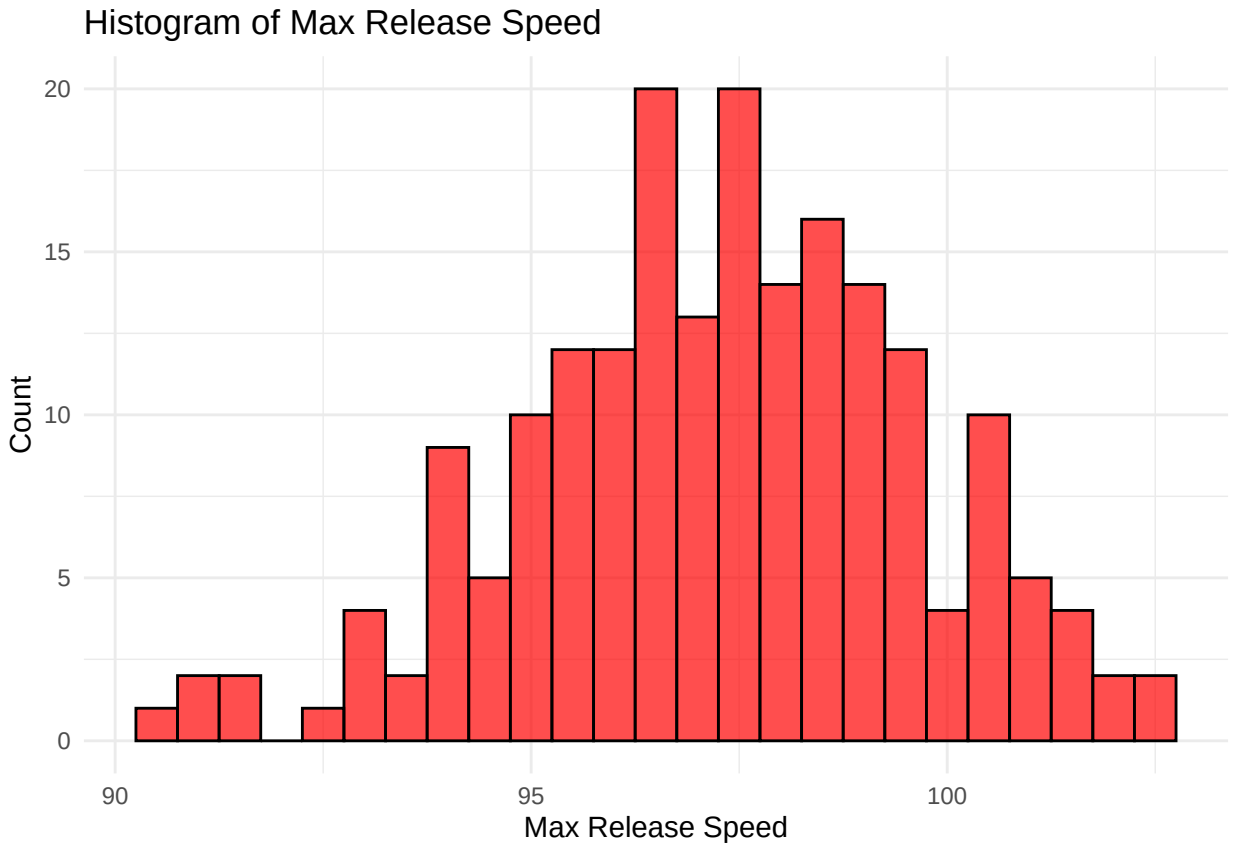
• line plot: Line plot to show the relation between two variables.

• qq-plot : These can be used to explain how close is the data to Normal distribution.

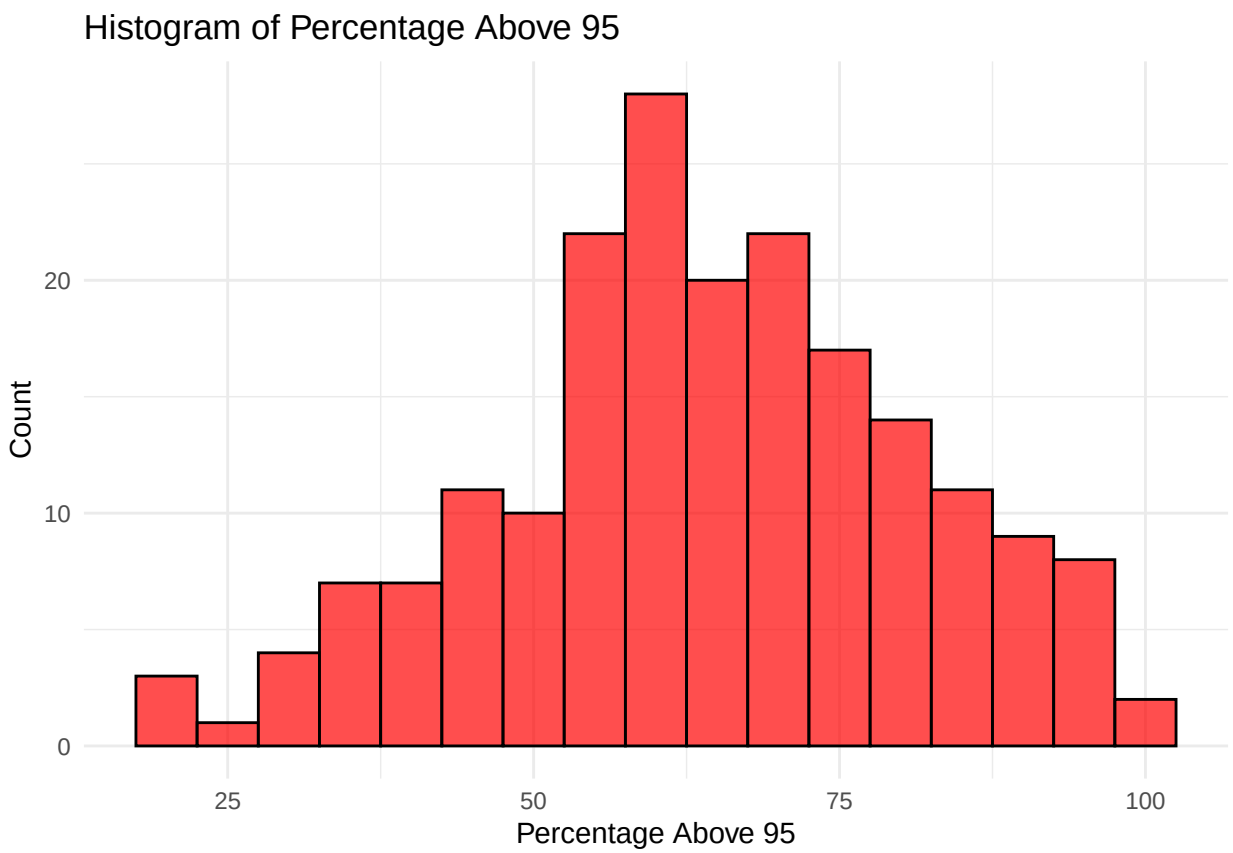
• Correlation Matrix: A table showing correlation coefficients between variables.

Histograms

```
# Histogram for max_release_speed
ggplot(final_pitch_data, aes(x = max_release_speed)) +
  geom_histogram(binwidth = .5, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Max Release Speed", x = "Max Release Speed", y = "Count") +
  theme_minimal()
```

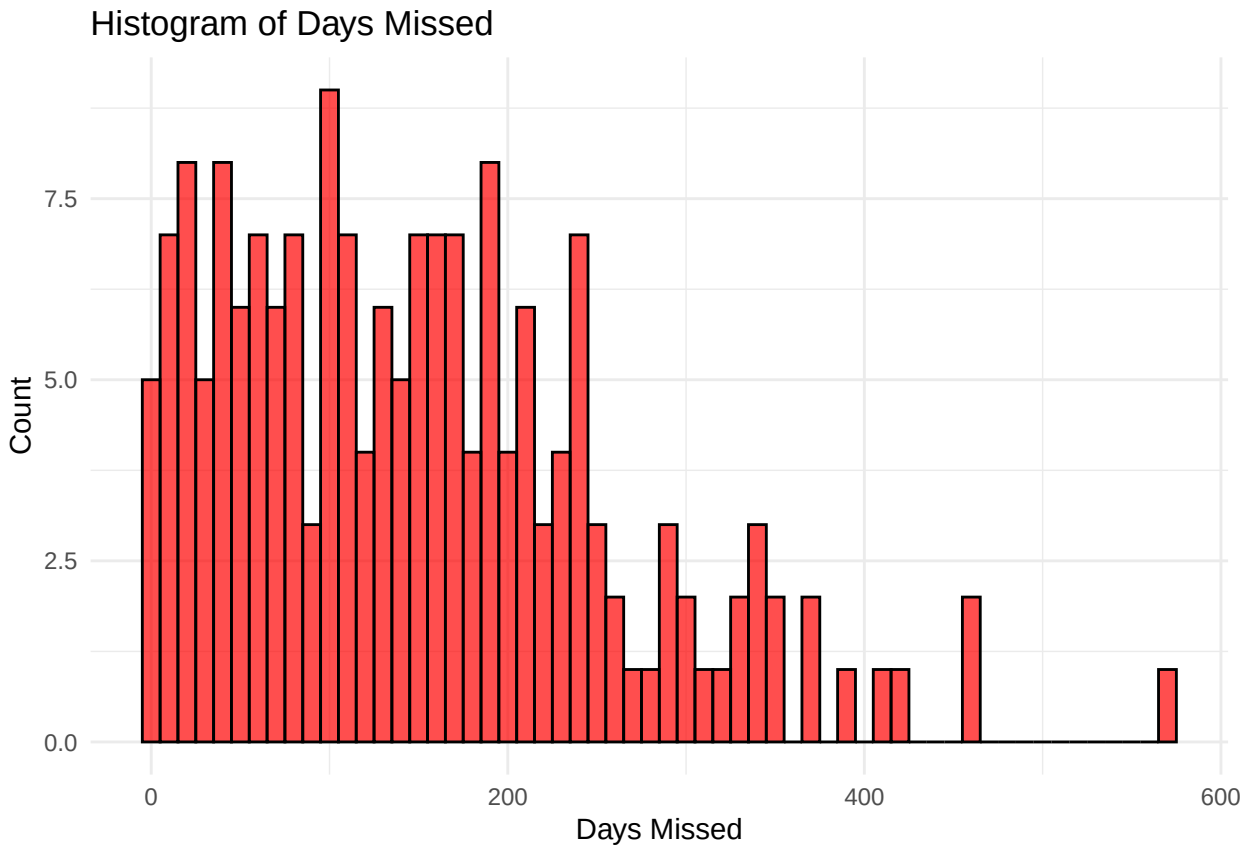



```
# Histogram for percentage_above_95
ggplot(final_pitch_data, aes(x = percentage_above_95)) +
  geom_histogram(binwidth = 5, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Percentage Above 95", x = "Percentage Above 95", y = "Count") +
  theme_minimal()
```



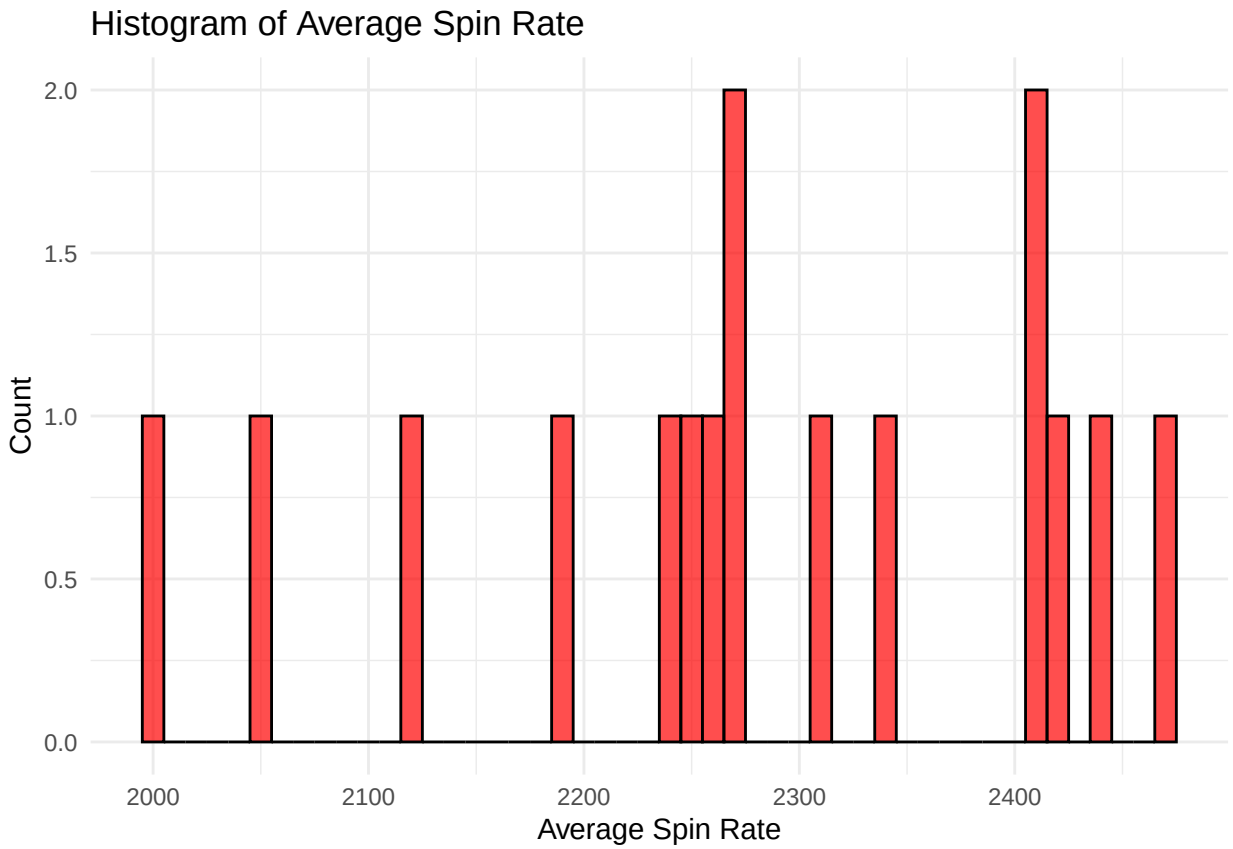

```
# Histogram for Days_missed
ggplot(final_pitch_data, aes(x = Days_missed)) +
  geom_histogram(binwidth = 10, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Days Missed", x = "Days Missed", y = "Count") +
  theme_minimal()
```

```
## Warning: Removed 17 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

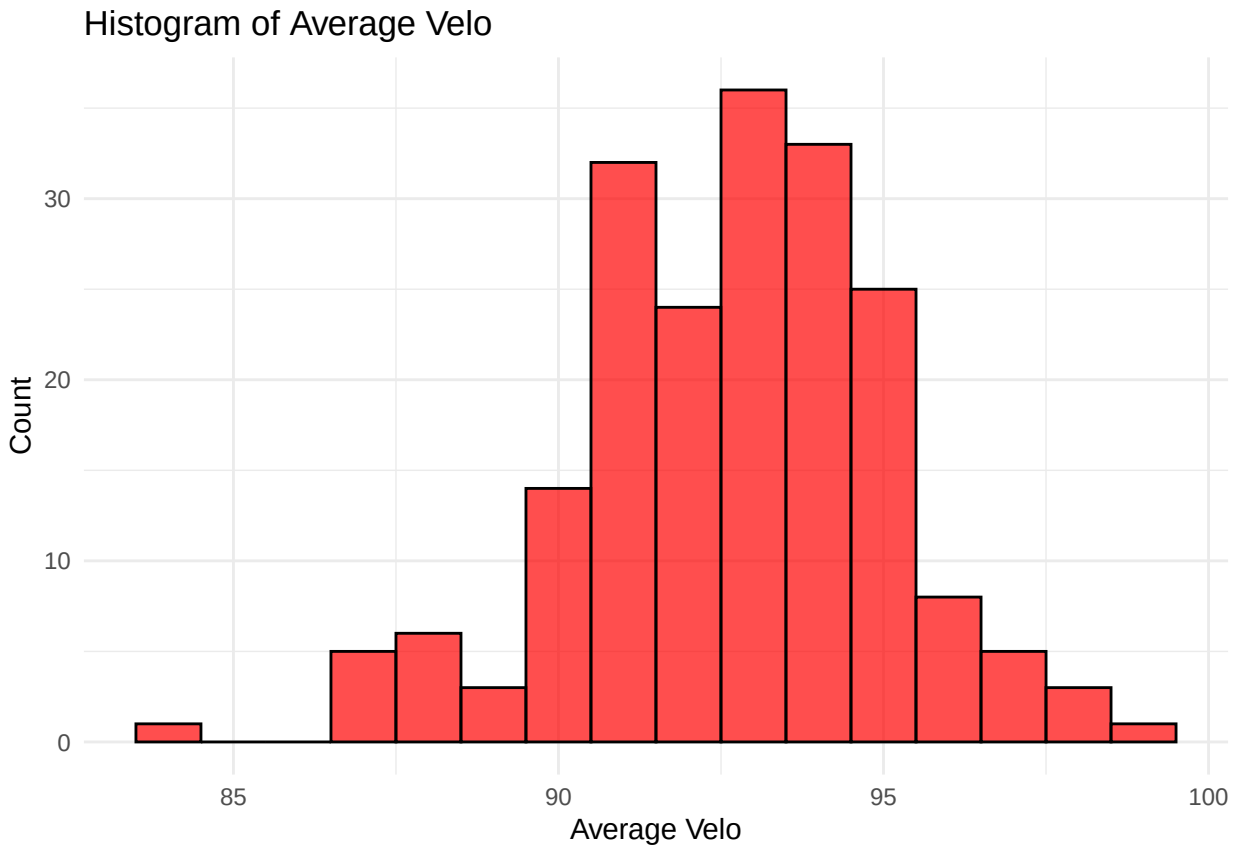



```
# Histogram for average_spin_rate
ggplot(final_pitch_data, aes(x = average_spin_rate)) +
  geom_histogram(binwidth = 10, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Average Spin Rate", x = "Average Spin Rate", y = "Count") +
  theme_minimal()
```

```
## Warning: Removed 180 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```r
# Histogram for average_velo
ggplot(final_pitch_data, aes(x = average_velo)) +
  geom_histogram(binwidth = 1, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Average Velo", x = "Average Velo", y = "Count") +
  theme_minimal()
```

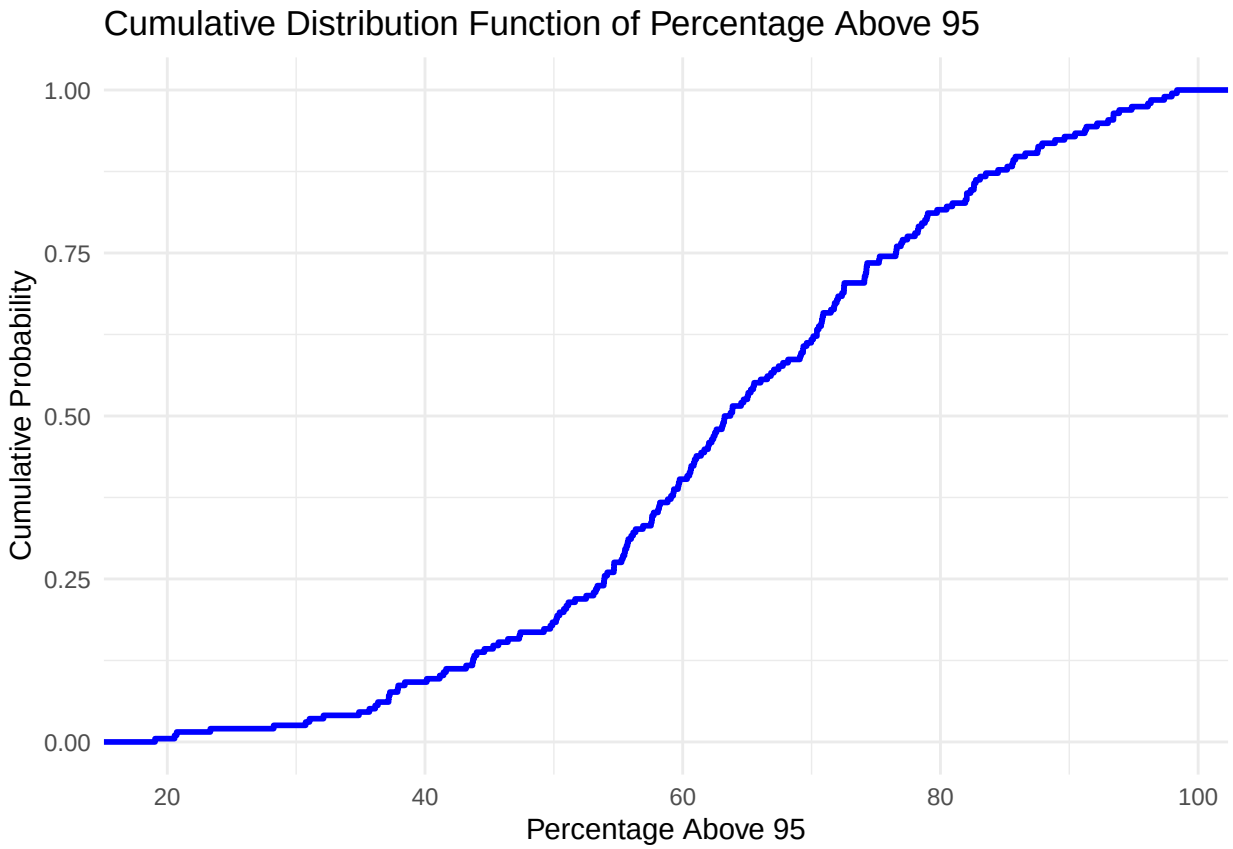


Observing the histograms of the quantitative variables reveals important information going forward. One thing that stands out is the number of holes in the "average_spin_rate" data. Glimpsing at the data table confirms that there are many NA values of this variable. This could be because spin rate is a relatively new measurement and there is limited spin rate data available going back five years. It could also be the case that equipment to measure spin rate is not available in all Major League ballparks. Whatever the reason for these missing values, this early analysis indicates that spin rate will not be a viable tool to define "effort" for the purpose of this research.

Another thing that stands out is the positive skewness of thee "Days_missed" variable. This is in part because of players who have not missed any time due to injury in the last five years, but there is also an extreme outlier of nearly 600 days missed. Removal of that outlier or a transformation of the data might be necessary for the most accurate modeling.

Cumulative Distribution Function

```
# Create the CDF plot
ggplot(final_pitch_data, aes(x = percentage_above_95)) +
  stat_ecdf(geom = "step", color = "blue", linewidth = 1) +
  labs(title = "Cumulative Distribution Function of Percentage Above 95",
       x = "Percentage Above 95",
       y = "Cumulative Probability") +
  theme_minimal()
```

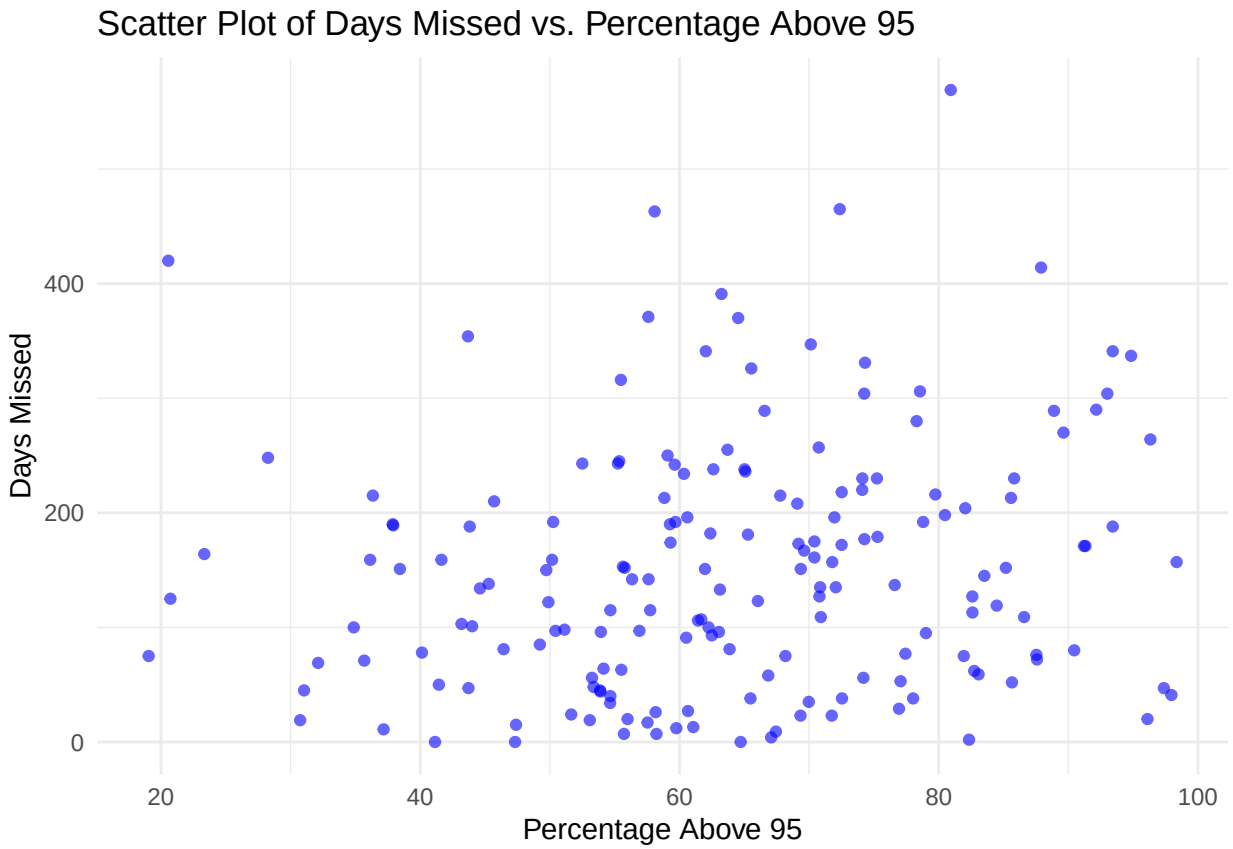


This CDF of the "percentage_above_95" variable indicates that about half of the observations fall below the value 65, and the other half above. This information will be valuable in hypothesis testing because I want to examine the mean number of Days Missed between two groups: pitchers who at or near their maximum velocity and those who attempt to throw their hardest less often.

Scatter Plot

```
ggplot(final_pitch_data, aes(x = percentage_above_95, y = Days_missed)) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(title = "Scatter Plot of Days Missed vs. Percentage Above 95",
       x = "Percentage Above 95",
       y = "Days Missed") +
  theme_minimal()
```

```
## Warning: Removed 17 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



There are no distinct patterns in the graph plotting "percentage_above_95" against "Days_missed". Pearson's correlation coefficient between the two variables is 0.161, indicating a very weak positive correlation. This indicates that a linear regression model will not be the best option for making predictions. Other models, such as k-nn clusters or random forests, may prove better model options, and the best model will likely also include additional variables.

Hypothesis Test

```
# Remove missing values in Days_missed
df <- final_pitch_data %>% filter(!is.na(Days_missed))

# Create the grouping variable
df$group <- ifelse(df$percentage_above_95 >= 70, "High", "Low")

# Perform an independent t-test
t_test_result <- t.test(Days_missed ~ group, data = df, var.equal = FALSE)

t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  Days_missed by group
## t = 2.3772, df = 128.1, p-value = 0.01892
## alternative hypothesis: true difference in means between group High and group Low is not equal to 0
## 95 percent confidence interval:
##   6.77022 73.99471
```

```
## sample estimates:
## mean in group High  mean in group Low
##           176.5075           136.1250
```

Using the CDF of "percentage_above_95" as a guide, I chose to test the mean Days Missed values of pitchers who throw 95% or more of the max velocity more than 70% of the time against those who exert this high amount of effort less than 70% of the time. The null hypothesis is that a higher "percentage_above_95" value will have no difference on the mean of days missed due to injury. The test produced a t-statistic of 2.377 and a p-value of 0.0189. A p-value of less than 0.05 suggests that we can reject the null hypothesis. There is a statistically significant difference in the mean of days missed due to injury between players with a "percentage_above_95" value above 70 and those below.

Implications

As indicated by the hypothesis test, there is a statistically significant difference in the means of time missed due to injuries between pitchers who throw very close to their maximum fastball velocity at least 70 percent of the time and those who more often throw at less than 95 percent of their max. The mean days missed for those in the higher "effort" group missed an average of 176.5 days over the five year span, while those in the lower "effort" group missed an average of 136.12 days. That difference of about 40 days over the five year span works out to an average of about eight extra days spent on the Injured List for those pitchers who like to consistently approach their maximum velocity when throwing a fastball.

Limitations

More research will be needed to determine if 95% of maximum velocity is the best value to test against injury. In addition to this measurement of "effort" other factors that might contribute to injury should be included in further analysis and model building. Those factors include, but are not limited to, bio-mechanics (arm angle at release, hip rotation, etc.), injury history, rule changes that might increase or decrease the pace of play, and environmental factors. Once the best variables for testing are selected, more research and testing will be required to determine the best model for accurate predictions.

Conclusion

Injuries are incredibly difficult to predict with even the most robust data available. What this study has done is provide a tool, or a piece of the puzzle, for the larger problem of predicting and reducing the time that a player might spend away from their respective team and on the Injured List. This analysis has shown a statistically significant relationship between throwing at or near a maximum velocity with a fastball at least 70 percent of the time and time spent on the Injured List. Determining how to best measure "effort" for pitches that are not fastballs is the next best step for this line of research.