Baseball Statistical Analysis

Average Velocity (MPH) vs. Strikeout Percentage (%)

By: Calvin Lee

1. Introduction

1.1 Client/Message Conveyed

The client this analysis would target would consist of primarily MLB/Baseball Scouts as various pitchers at every age can throw at many different velocities. As well as this, my client would further expand to any baseball scouts in the world and any keen baseball fans attempting to get a better insight into how the modern game of baseball functions. The message being conveyed consists of the correlation between how the average velocity thrown causes higher strikeout percentages, overall showing the importance of acquiring pitchers that throw harder as the MLB progresses towards the future.

1.2 Personal Background/Reason study was chosen

Overall this topic was chosen due to baseball being a field of interest of mine from a young age. Starting to play at the young age of 9 baseball has been a fundamental sport that I have grown up and loved to watch/play. Furthermore, this field of study was chosen due to the sheer high quantity of accessible data produced by the MLB, especially for pitchers. With an average of 8 Games per day league-wide from spring to fall and an average of 288 pitches thrown each day, this creates 2304 new data points to be created each day, allowing for more accurate and insightful research around this topic.

1.3 Importance of this study

Determining the future stars in the MLB shapes the future state of baseball. By analyzing the Average Velocity thrown by a pitcher vs its respective strikeout percentage, This analysis could revolutionize the future of Baseball and scouting, aiding to create
Future scouting methods to develop the techniques to find Top Talent. It is intuitively known that throwing harder tends to lead to more strikeouts. But how strong is this relationship? Is it a straight line, or does it plateau at a certain velocity? This analysis could quantify and help us to understand the true impact of velocity.

1.4 Previous research is done in this Field

According to Dr. Josh Heenan, there is currently an average of less than 10 pitchers that can consistently throw over 100 MPH. Researching the impacts of high-velocity pitching could cause a switch in priorities for many pitchers around the world revolutionizing the game of baseball and the future of pitching.

2. Data Collection Methodaligies

2.1 Sources

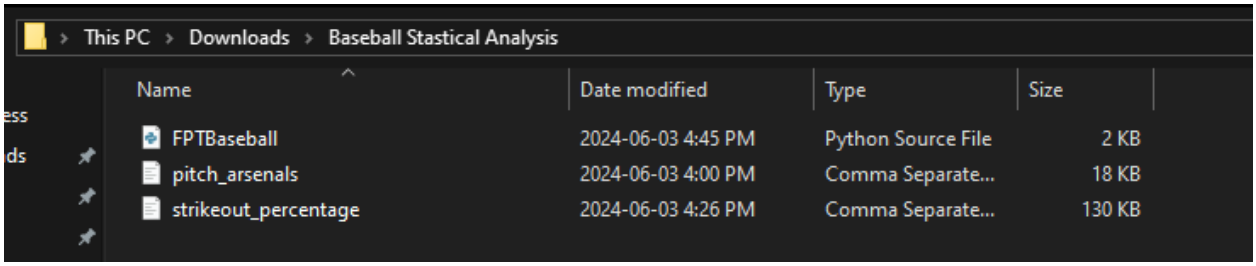
BaseballSavant and Baseball-reference will be used to gather the data for this statistical study, which will include the entirety of the MLB 2023 regular season.

The average 4-seam fastball velocity will be used as the key mark to compare velocities across MLB pitchers, provided by a dataset found on Baseball Savant. In the case an MLB pitcher doesn't throw a 4-seam fastball, the average velocity of that pitcher's sinker will be utilized in its instead. Furthermore, the average strikeout percentage will be used to compare the average strikeout rate between pitchers in the 2023 MLB season, provided by the dataset provided on the website Baseball-reference.

Both of these respective websites and datasets have been chosen due to their availability to be downloaded as a CSV file, to be then compared through the techniques of data parsing within Python’s File-IO system.

2.2 Comparing the data

As previously mentioned, this data will be compared using Python and techniques of data parsing using the File-Io system. Initially, both CSV files containing both variables were uploaded into a folder with a new py file.



Following this, a Python script was written to parse the values of the average velocity and strikeout percentage together Using the pitcher's name as a comparison value to equate the two pieces of data together. Both of these data pieces were put together using a nested dictionary in the form

{ "Pitcher's Name" : ["Average pitching Velocity" , "Average Strikeout Percentage"], …}

```python
import csv
filename1 = "pitch_arsenals.csv"
filename2 = "strikeout_percentage.csv"

database = {}

with open(filename1, 'r') as file_in:
    file_in.readline()
    csv_reader = csv.reader(file_in)

    for line in csv_reader:
        unparsed_name = line[0]
        name_list = unparsed_name.split()
        full_name = name_list[1] + ' ' + name_list[0][:-1]

        fastball_speed = line[2]
        sinker_speed = line[3]

        if fastball_speed == "":
            fastball_speed = sinker_speed

        try:
            database[full_name] = [fastball_speed]
        except KeyError:
            pass
```

```
26
27
28  with open(filename2, 'r') as file_in:
29      file_in.readline()
30      csv_reader = csv.reader(file_in)
31      for line in csv_reader:
32          name = line[1]
33          if name[-1] == "*":
34              name = name[:-1]
35
36          so_percent = line[10]
37
38
39          try:
40              database[name].append(so_percent)
41          except KeyError:
42              pass
43
```

Once the dictionary was compiled, a CSV file was written using the provided dictionary writing each line in the form [Average_Velocity, Strikeout_Percentage], to be then opened using spreadsheets for graphs to then be produced to visualize the statistical findings.

```
43
44  with open('baseball_data.csv' , 'w') as file_out:
45      csv_writer = csv.writer(file_out)
46      write_lines = []
47      for i in database:
48          write_line = []
49          if len(database[i]) < 2:
50              pass
51          else:
52              pitch_velocity = database[i][0]
53              strikeout_percent = database[i][1]
54
55              temp_line = [pitch_velocity, strikeout_percent]
56              write_lines.append(temp_line)
57      csv_writer.writerows(write_lines)
58
```

C: > Users > Calvin > Downloads >

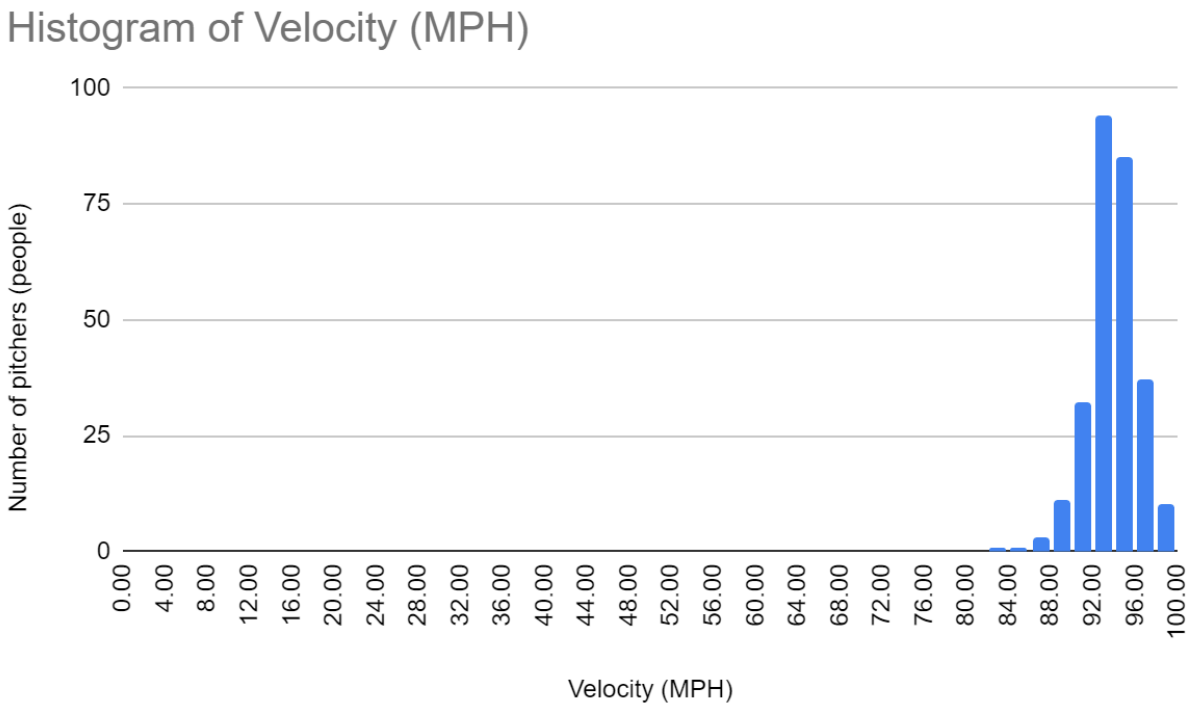
```
1   96.7,27.0%
2   95.6,27.3%
3   93.6,26.0%
4   96.3,27.3%
5   93.3,15.9%
6   93.1,25.7%
7   92.6,22.8%
8   95.5,31.5%
9   92.4,23.6%
10  95.8,26.9%
11  92.4,22.5%
12  95.2,25.5%
13  97.2,36.8%
14  92.7,25.5%
```

With this complete, a one and two-variable statistical analysis could then be conducted to compare the correlation between Average Velocity thrown vs Strikeout Percentage.

3. One-Variable Statistics

3.1 Velocity (MPH)

Using functions within Google Sheets, the following variables and histograms were calculated and created.



Analysis	
	Velocity (MPH)
n	274
max	99.5
min	82.8
range	16.7
mean	93.81350365
median	93.85
mode	95.8
Q1	92.525
Q2	93.85
Q3	95.5
Interquartile Range	2.975
Standard Deviation	2.39934365

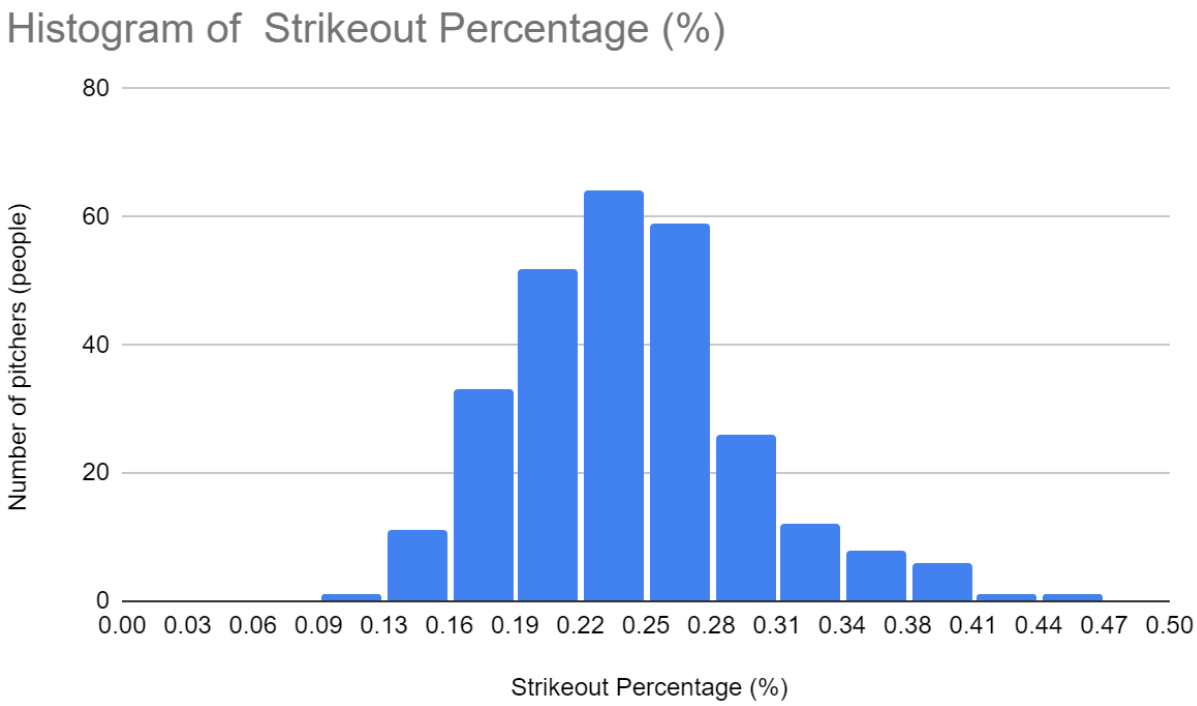
It was determined that the Best measure of central tendency is the Mean, due to strong deviations in this graph demonstrating high or low-skill ability players, and those people being a key part and importance to baseball and should therefore be accounted for in the measure of spread.

Furthermore using the formulas
Low Outlier < Q1 - 1.5(IQR) and High Outlier > Q3 + 1.5(IQR) it was determined that there are a total of 5 low outliers in this dataset.

IQR: 2.975 Low Outlier < Q1 - 1.5(IQR) Low Outlier < 92.525 - 1.5(2.975) Low Outlier < 88.06 There are 5 low outliers	High Outlier > Q3 + 1.5(IQR) High Outlier > 95.5 + 1.5(2.975) High Outlier > 99.96 There are no high outliers

3.2 Strikeout Percentage (%)

Using functions within Google Sheets, the following variables and histograms were calculated and created.



Analysis	
	Strikeout Percentage (%)
n	274
max	0.46
min	0.11
range	0.35
mean	0.24
median	0.24
mode	0.214
Q1	0.2
Q2	0.237
Q3	0.27175
Interquartile Range	0.07175
Standard Deviation	0.05778849977

It was determined that the Best measure of central tendency is the Mean, due to strong deviations in this graph demonstrating high or low-skill ability players, and those people being a key part and importance to baseball and should therefore be accounted for in the measure of spread.

Furthermore using the formulas
Low Outlier < Q1 - 1.5(IQR) and High Outlier > Q3 + 1.5(IQR) it was determined that there are a total of 9 high outliers in this dataset.

IQR: 0.07175 Low Outlier < Q1 - 1.5(IQR) Low Outlier < 0.2 - 1.5(0.07175) Low Outlier < 0.092’ There are no Low outliers	High Outlier > Q3 + 1.5(IQR) High Outlier > 0.27175 + 1.5(0.07175) High Outlier > 0.3794 There are 9 High Outliers in this data

4. Two-Variable Statistics

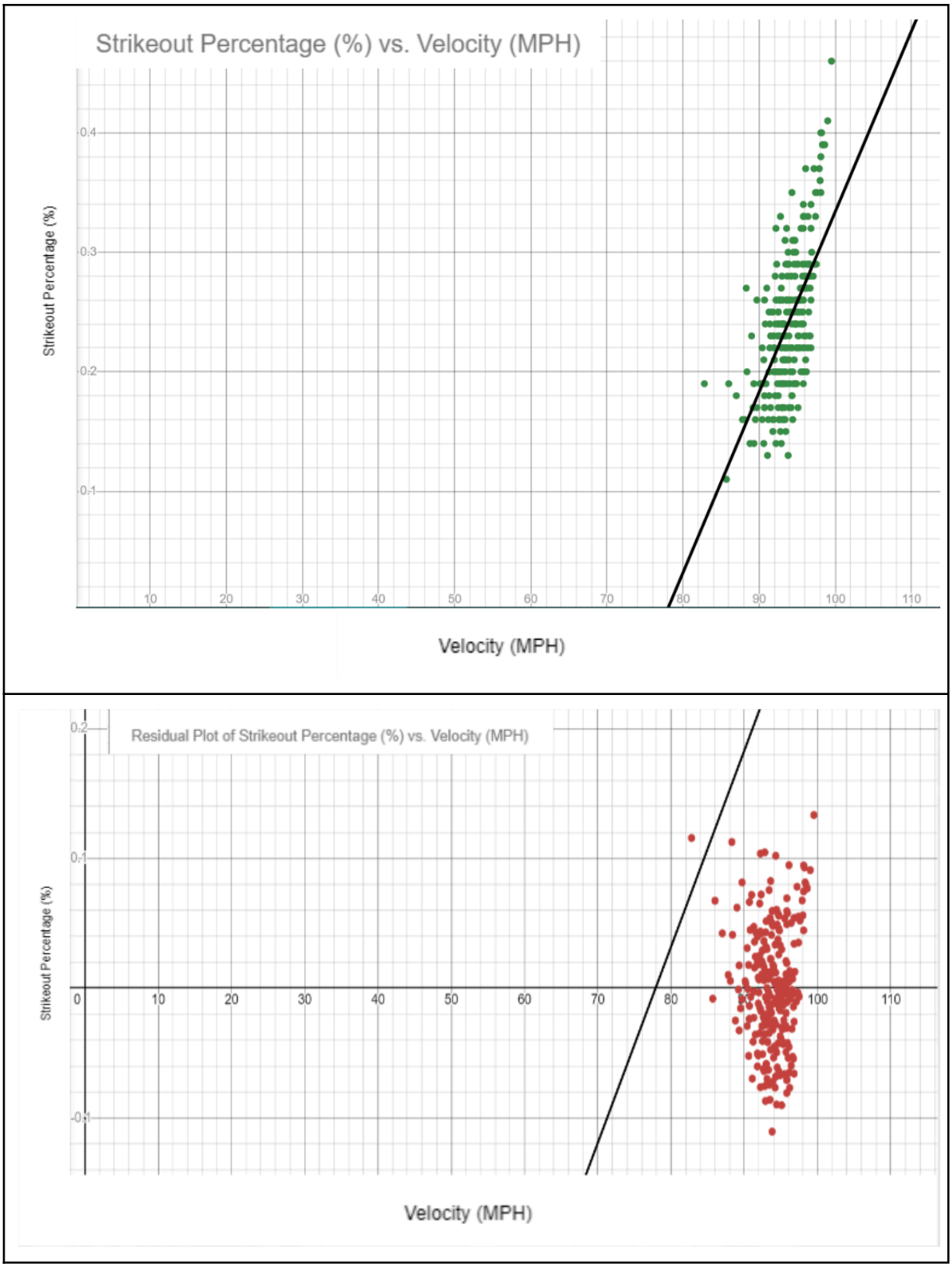
4.0 Models

There are a few models to consider for when determining regression between these two variables.

4.1 Linear Regression

Using values compiled in a spreadsheet, the following linear regression and residual plot graphs were constructed.

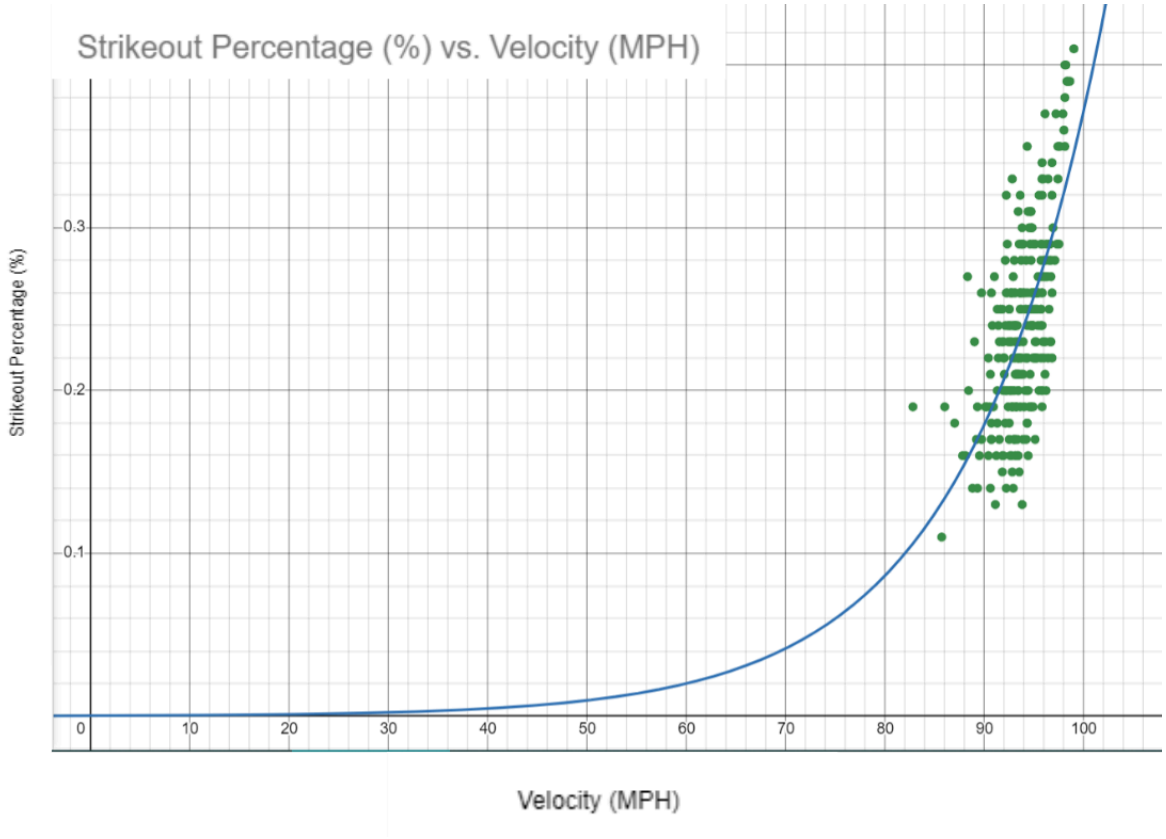
Formula: y = 0.0151096x - 1.17683
R = 0.6245 **(Moderate Positive Correlation)**
R^2 = 0.3901

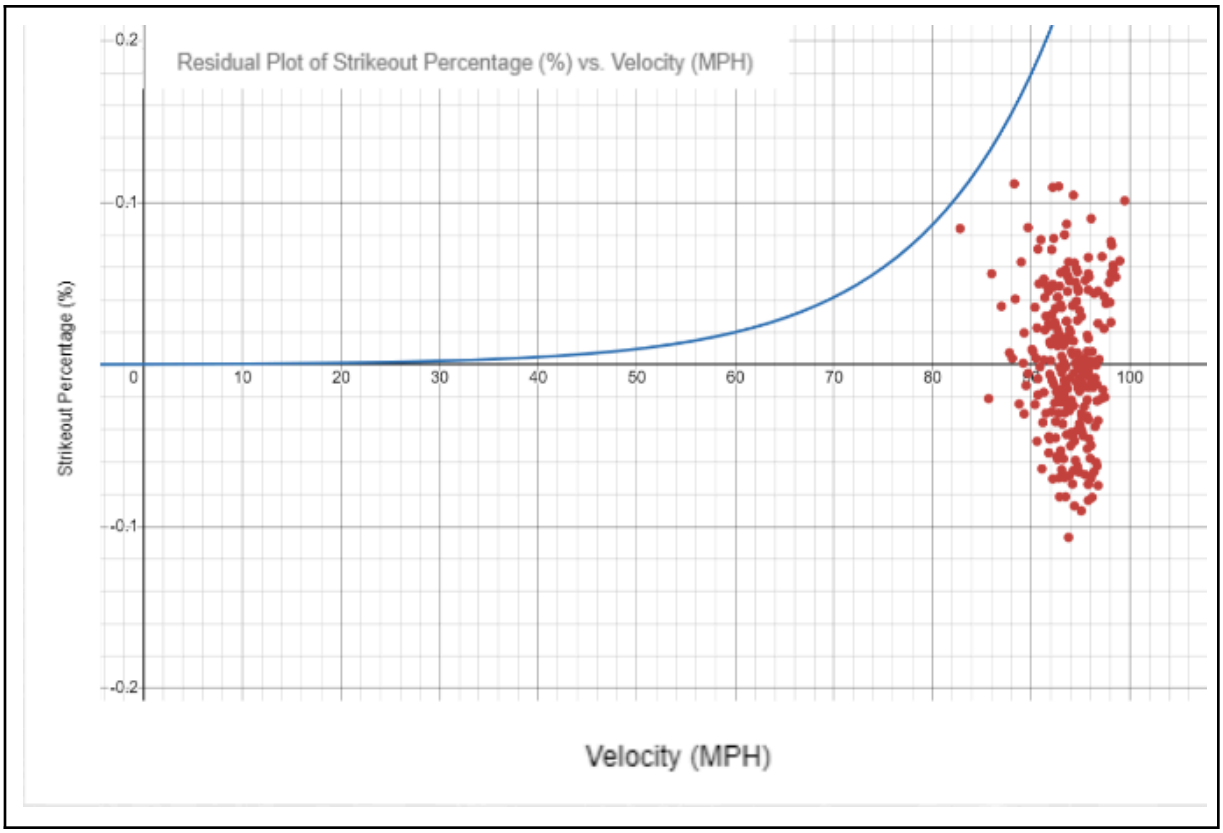

Strikeout Percentage (%) vs. Velocity (MPH)
Strikeout Percentage (%)
0.4
0.3
0.2
0.1
10
20
30
40
50
60
70
80
90
100
110
Velocity (MPH)
Residual Plot of Strikeout Percentage (%) vs. Velocity (MPH)
Strikeout Percentage (%)
0.2
0.1
0
-0.1
10
20
30
40
50
60
70
80
100
110
Velocity (MPH)

4.2 Exponential Regression

Using values compiled in a spreadsheet, the following exponential regression and residual plot graphs were constructed.

Formula: $y = 0.000250736(1.07576)^X$
$R^2=0.436$





4.3 Better model for the dataset

Overall the Exponential model best fits the dataset. This is due to a couple of reasons

R^2 value:

Linear regression R^2 value : $R^2 = 0.3901$
Exponential Regression R^2 Value: $R^2=0.436$

Better fit for baseball:

Overall the Exponential regression fits the data better evident from its base trajectory. In the linear regression, there features a negative y-int, creating negative strikeout percentages possible when throwing between 0-77.89 MPH.

4.4 Interpolation

To prove the Exponential Regression is a better fit for baseball, Let's say we want to find the strikeout percentage of a pitcher that throws an average speed of 50 MPH according to this data.

Linear:	**Exponential:**
y = 0.0151096x - 1.17683 y = 0.0151096(50) - 1.17683 y= -0.42 y= = -42%	y = 0.000250736(1.07576)X y = 0.000250736(1.07576)50 y= 0.0096 y= 0.96%

Since the linear Regression features a negative Strikeout percentage and the exponential doesn't. Therefore the exponential regression will fit better in this dataset.

5. Conclusion

5.1 Relationships

Looking at an overview of the data there is an approximate moderate positive correlation between the variables creating a moderate Casual relationship

As well as that, there is an Extraneous relationship with external variables such as command, decision-making, and luck impacting the overall strikeout percentage of a given pitcher

5.2 Fault with Dataset

Overall one major factor unconsidered when producing this analysis was that All data comes from MLB players.

The average high school baseball player has a 0.05% chance of making the MLB, demonstrating how the data used for these findings does not accurately represent the entirety of the baseball pitching population.

These effects were evident in this comparison, as the projected values within the Average velocity One variable statistics,

showed the world's top baseball players, causing little deviation in the data and an inaccurate representation of baseball's whole field. As a result, the data was skewed and compressed.

5.3 Overall Findings

Overall in conclusion, There is a positive correlation between Velocity (mph) vs Strikeout Percentage (%), although there may be other factors that impact strikeout percentage such as command and pitching decision.

Citations

"2023 Major League Baseball Advanced Pitching." *Baseball*, www.baseball-reference.com/leagues/majors/2023-advanced-pitching.shtml.

"Statcast Pitch Arsenals Leaderboard." *Baseballsavant.Com*, baseballsavant.mlb.com/leaderboard/pitch-arsenals.

Shields, Aaron. "The Odds of Making It to MLB." Casino.Org Blog, 12 Feb. 2024,www.casino.org/blog/odds-of-making-it-to-mlb/#:~:text=But%20overall%2C%20that%20means%20that,drafted%20by%20an%20MLB%20team

Cooper, JJ. "Ever-Climbing Velocity Pushes Hitters to the Brink" Baseball Americahttps://joshheenan.com/what-does-a-100mph-fastball-look-like-in-the-gym/#:~:text=Throwing%20100MPH%20is%20the%20goal,pitchers%20touch%20100MPH%20per%20season.