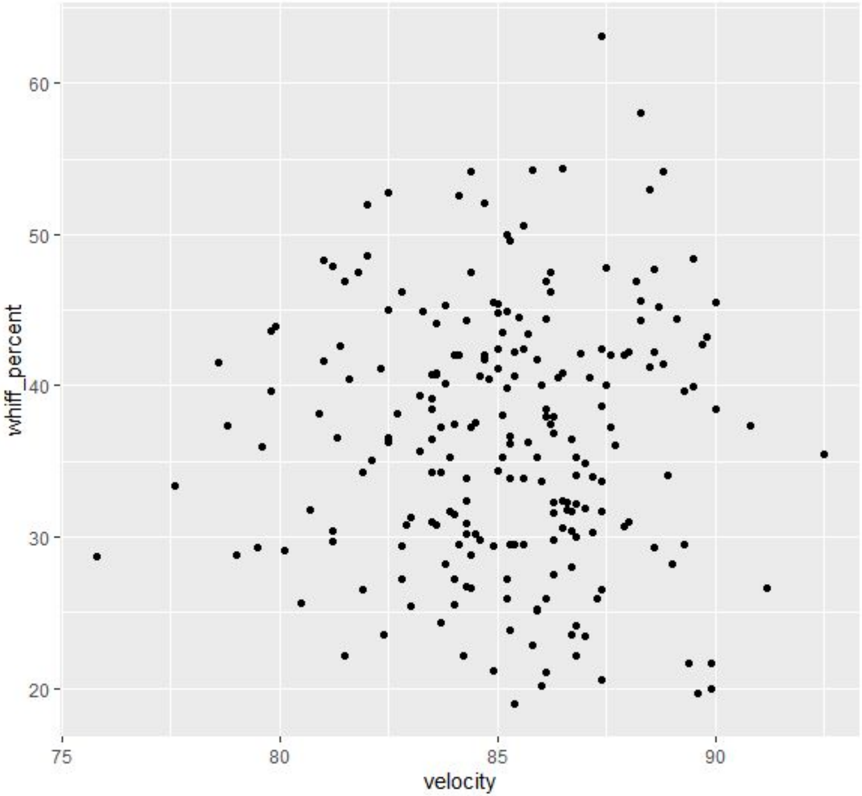
Optimal Pitch: Spin Rate & Velocity to Maximize Whiff Rate

By: Minsoo Park, Richard Yang,
Neil Rowe, and Wesley Fletcher

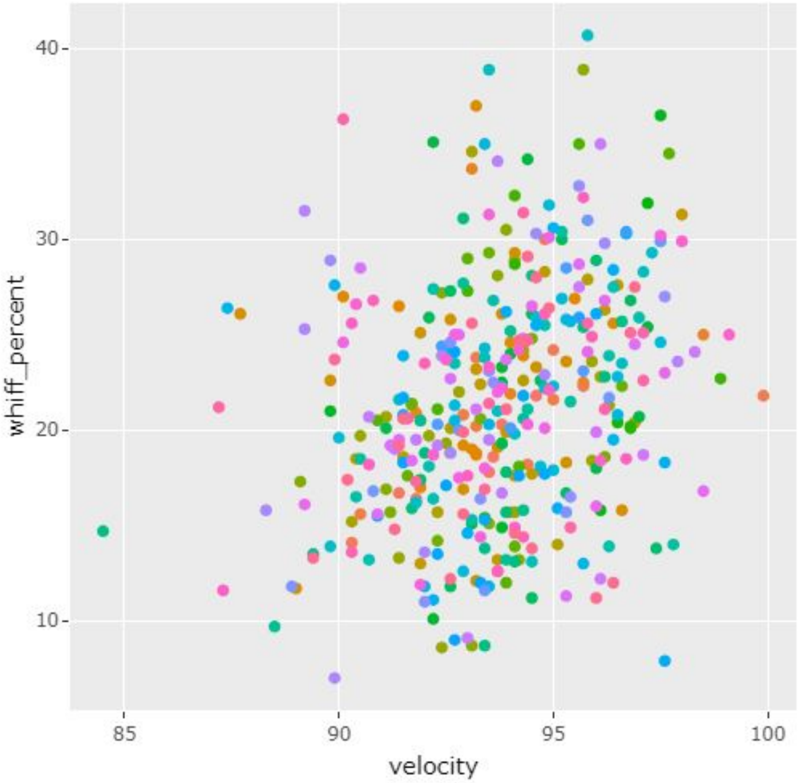
Research Hypothesis: Spin rate and velocity will affect the whiff rate

Null Hypothesis: Spin rate and velocity have no correlation with whiff rate and the data is because of chance

Background



60
50
40
30
20
75
80
85
90
velocity
whiff_percent



40
30
20
10
85
90
95
100
velocity
whiff_percent

More Background

- Common Belief: Throwing Harder leads to whiffs
  - Not much correlation between velocity and whiff rate
    - r (4 seamers vs. whiff rate)=.273
    - r (Sinkers vs. whiff rate)=.323
    - r (Sliders vs. whiff rate)=.0079
    - r (Changeups vs. whiff rate)=.0766
  - Examples
    - Andres Munoz has on average the fastest fastball in the majors, but a pedestrian whiff rate
    - Tyler Clippard barely averages 90 mph on his fastball, but has one of the best whiff rates in baseball

If more velocity doesn't lead to more whiffs, then is spin the key factor that affects a pitcher's ability to miss bats?

Pitcher 1: Mike Minor

- Low Velo, High spin rate
  - For all of his pitches, his average velocity clocked in at around **86.93 mph**
  - However, his average spin rate maintained a high **2502.36 rpm**
  - Among the players we used in our data, Minor had the **5th** highest fastball spin rate, **8th** highest changeup spin rate, and **29th** highest slider spin rate

Pitcher 2: Nathan Eovaldi

- High velo, Low spin rate
  - Four-seam fastball average of **97.5 mph** in 2019
  - Cutter averaged **93.2 mph**
  - However, his fastball spin rate was below average
    with **2186 rpm**
  - His curveball spin rate was well below average at **2174 rpm**

Inquiry

Q: Any other factors that impact whiff rate? And if any, to what extent?

- Velocity
- Spin rate
- “Combination” of pitch repertoire

Univariate Regression showed not much correlation, but....

We still need to perform multivariate regression!

Basic stat info.

Mean: ΣX_i / N (μ)

Standard deviation: $\{\Sigma(X_i - \text{Mean})^2 / N\}^{0.5}$ (σ)

Variance: (Standard deviation)^2 (σ^2)

RMSE: $\{\Sigma(y_i - \hat{y}_i)^2 / N\}^{0.5}$ ($\hat{y}$ = predicted value of y)

R-squared: $1 - (SS_{Regression} / SS_{Total})$ where $SS_{Regression} = \Sigma(y_i - \hat{y}_i)^2$ and $SS_{Total} = \Sigma(y_i - \bar{y}_i)^2$ ($\bar{y}$ = mean of y)

Z-score: (X-mean) / S.D.

Univariate Linear Regression Model

Univariate Linear Regression Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameter: (θ_0, θ_1)

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$

Aim/Goal: to minimize cost $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

*** h(x) = Ŷ, $\boldsymbol{\theta}_n$ = Parameters, x_n = Features (only 1 feature), $y^{(i)}$ = Actual Output

Estimate unknown parameters for given x

Multivariate Regression Model

Exact same process as univariate linear regression, but with multiple features

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n$

$$h_\Theta(x) = \Theta^T x \quad \textit{where } \Theta = \sum \theta \quad \textit{and } T \textit{ means transposed matrix}$$

Parameter: $\theta_0, \theta_1, \theta_2, \ldots, \theta_n$

Cost Function: $J(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\Theta\left(x^{(i)}\right) - y^{(i)} \right)^2$

Aim/Goal: to minimize cost $\min_{\Theta} J(\Theta)$

Step 1: Data crawling process

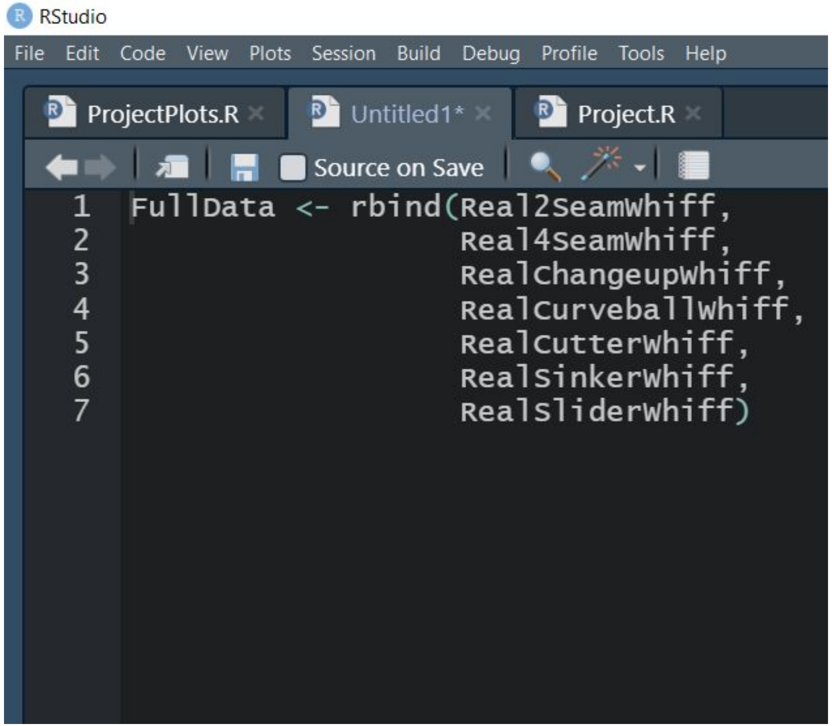
- Data gathered from statcast

- Crawled data that was in table format by converting it into .csv files

```
Real2SeamWhiff <- read.csv("data/Real2SeamWhiff.csv")

Real4SeamWhiff <- read.csv("data/Real4SeamWhiff.csv")

RealChangeupWhiff <-read.csv("data/RealChangeupWhiff.csv")

RealCurveballWhiff <- read.csv("data/RealCurveballWhiff.csv")

RealCutterWhiff <- read.csv("data/RealCutterWhiff.csv")

RealSinkerWhiff <- read.csv("data/RealSinkerWhiff.csv")

RealSliderWhiff <- read.csv("data/RealSliderWhiff.csv")
```

Step 2: Merging Data Frames

After reading all of the .csv files for each pitch type, we merged all 7 data frames into 1 master dataframe using the function "rbind"



Step 3: Mean Normalization (Z-Scores)

In regression, it is better to keep the values of all features within certain boundary (ex: between -1 and 1). But since artificially altering features is not recommended, decided to use feature scaling: Mean normalization method

Mean normalization: $x_{i,\, scaled} = \dfrac{x_i - \mu_i}{S_i}$

x_i = Data, μ_i = Average (mean) of population, S_i = S.D.

For our dataset (FullData):

- Velo mean: **89.19 mph** / Velo S.D.: **5.47 mph**
- Spin rate mean: **2262.22 rpm** / Spin rate S.D.: **284.83 rpm**

Step 4: Filtering

In order to remove outliers from our data that may skew our graph, we constrained our data points to exclude points that we found to be way too extreme.

We repeated this for all the types of pitches by using the filter() function on our dataframe

```
FullData2 <-
  FullData %>%
  mutate(pred = my.lm$fitted.values) %>%
  filter(pitch_type %in% c("Cutter"))

x1 <- FullData2$spin_rate
x2 <- FullData2$velocity
y <- FullData2$whiff_percent
dataset <- cbind.data.frame(x1,x2,y)
scatterplot3d(x1,x2,y)
```

Example: Filtering by Changeup Pitches

Step 5: Regression

```
FullData <-
  FullData %>%
  mutate(whiff_percent = whiff_percent/100)

baseballpitcher <- lm(whiff_percent ~ spin_rate + pitch_type, data = FullData)
summary(baseballpitcher)
```

Since we are dealing with *both* spin rate and velocity....

Using the linear model “lm()” function, we found the summary of our multivariable regression, which showed that....

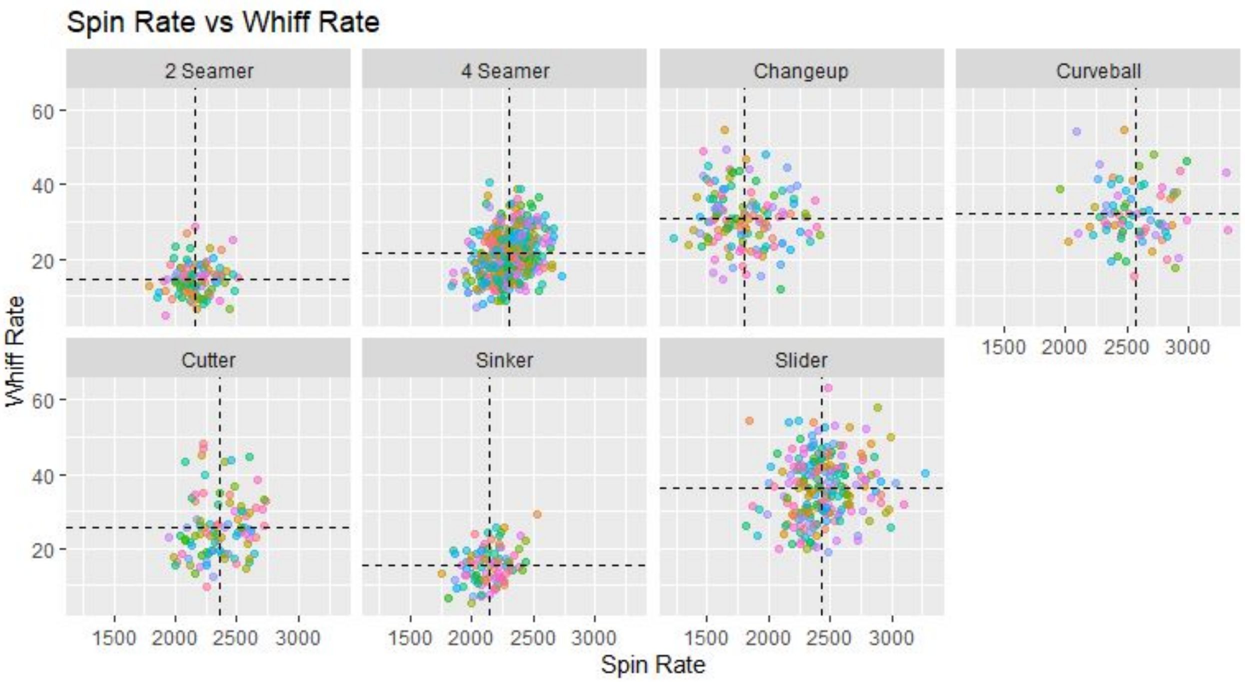
Multiple Regression Findings

- The p-value was extremely low, a sign that our findings were in fact **statistically significant**
- The correlation was pretty high compared to the values we observed earlier with the univariate regression model

r = 0.5012

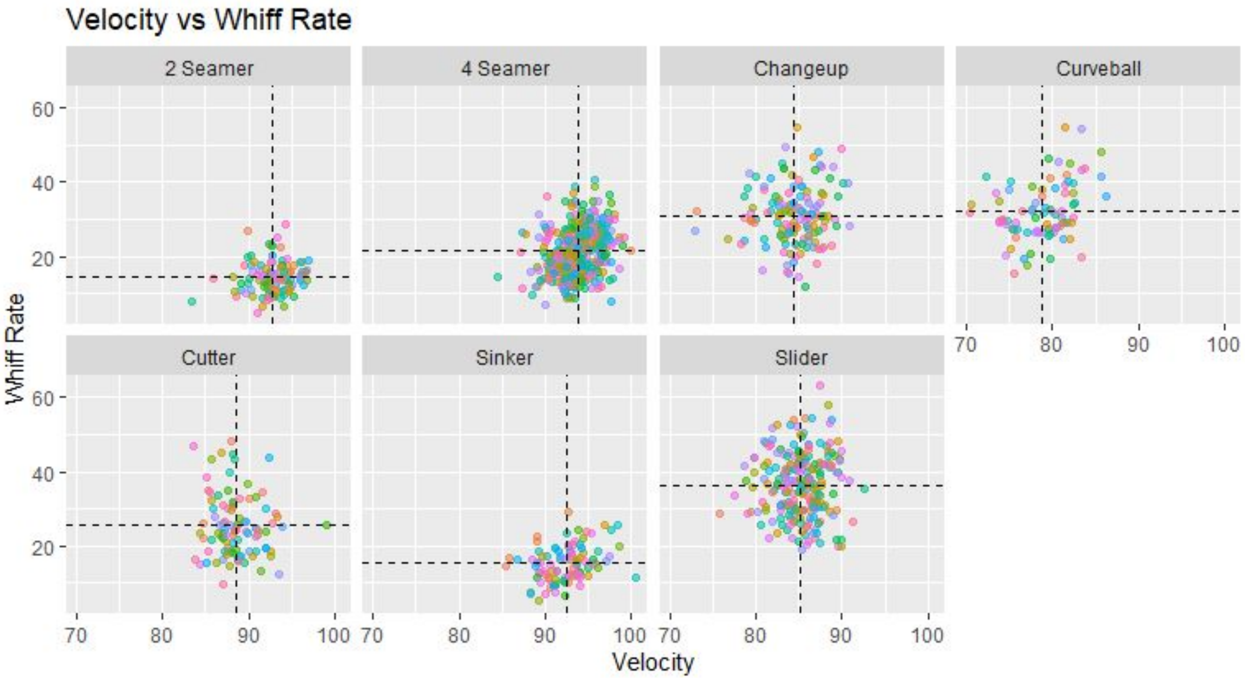
```
Call:
lm(formula = whiff_percent ~ spin_rate + velocity, data = FullData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.24201 -0.06357 -0.00800  0.05665  0.34656 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.278e-01  4.859e-02  19.097  < 2e-16 ***
spin_rate    5.858e-05  9.264e-06   6.323  3.7e-10 ***
velocity    -9.026e-03  4.826e-04 -18.702  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0885 on 1124 degrees of freedom
Multiple R-squared:  0.2616,	Adjusted R-squared:  0.2603 
F-statistic: 199.1 on 2 and 1124 DF,  p-value: < 2.2e-16
```

Plotting Data



Using the facet_wrap() function with the ggplot() function, we created graphs for each pitch type comparing spin rate and whiff rate

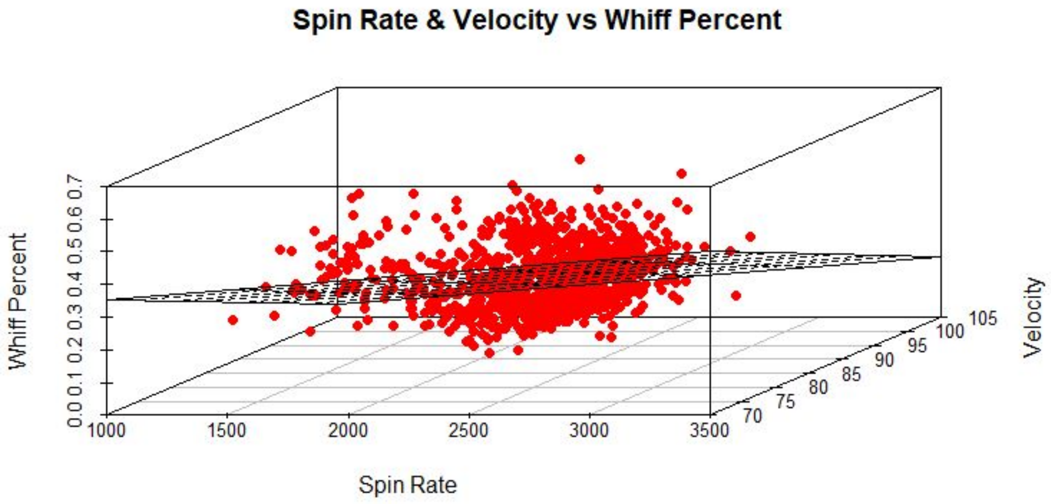
Plotting Data Cont.



In addition, we added the mean lines for the x-axis (pitch stat) and y-axis (whiff rate) with the geom_vline() and geom_hline() functions

Multivariate Plots For Each Pitch Type (3D)

All Pitches



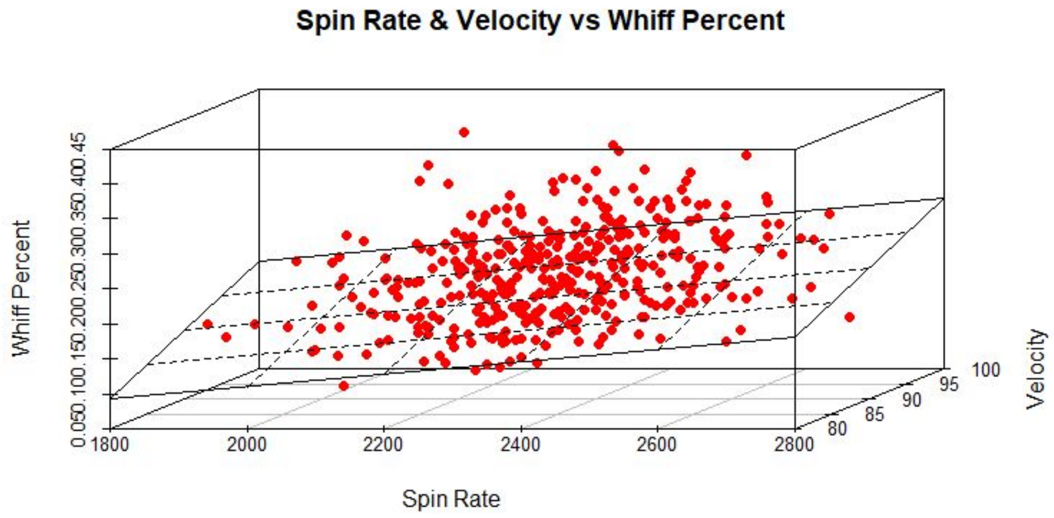
Trend:

- As **Spin Rate** goes up, so will the whiff rate
- As **Velocity** goes up, the whiff rate will go down

Optimization: High Spin, Low Velocity (but probably due to breaking balls having lower velo and higher whiff rates => **"Simpson's Paradox"**)

3D Plots

4 Seamer



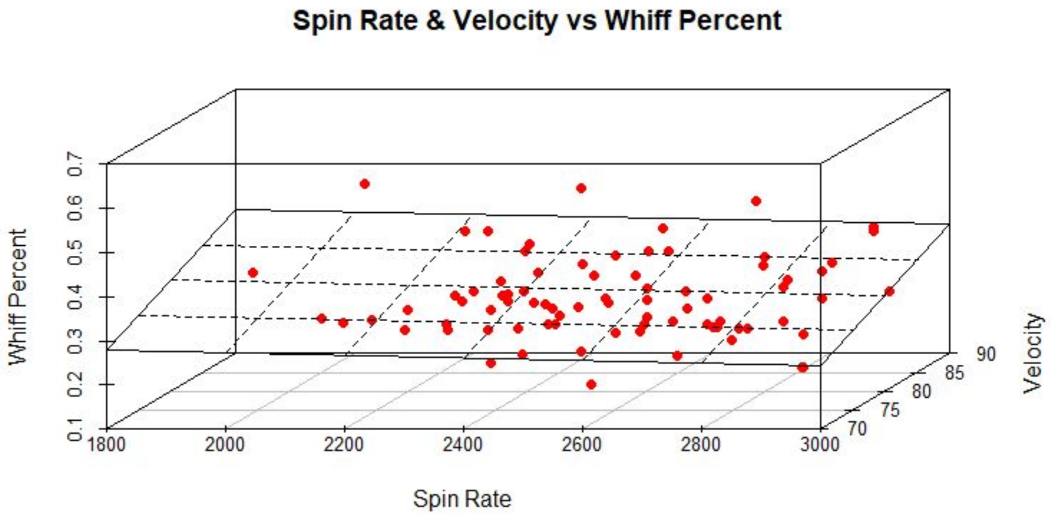
Trend:

- As **Spin Rate** goes up, the whiff rate increases
- As **Velocity** goes up, the whiff rate increases

Optimization: (+, +)

3D Plots

Curveball



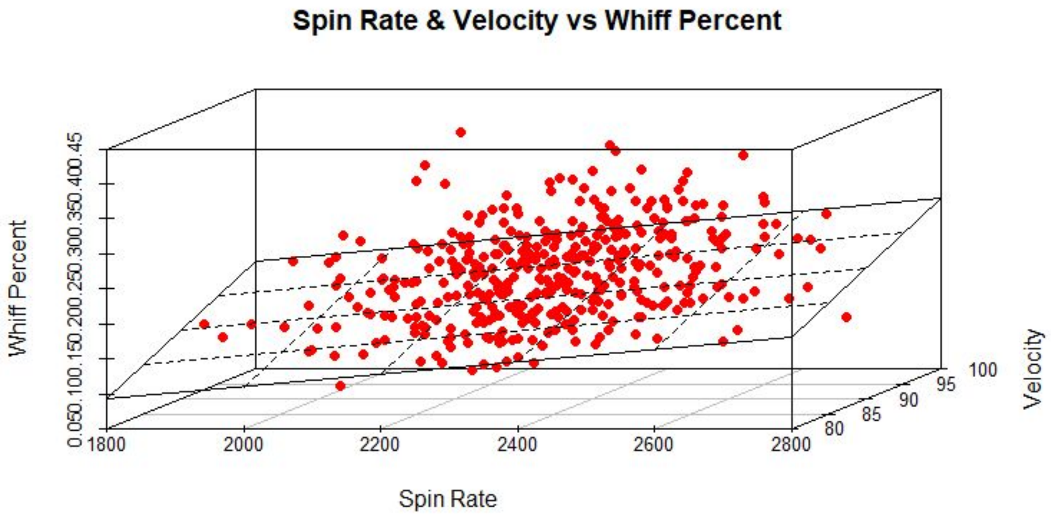
Trend:

- As **Spin Rate** goes up, the whiff rate stays constant
- As **Velocity** goes up, the whiff rate increases

Optimization: (null, +)

3D Plots

Changeup



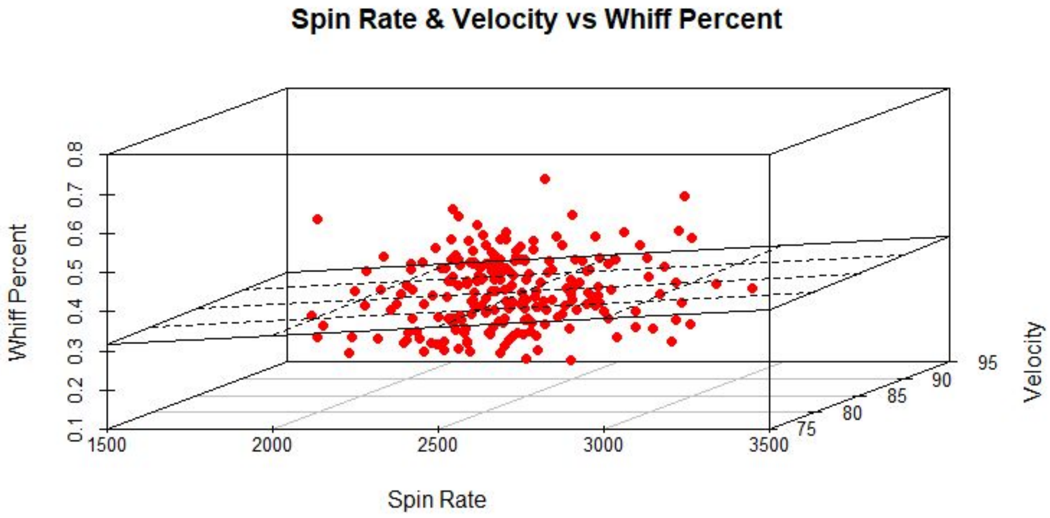
Trend:

- As **Spin Rate** goes up, the whiff rate increases
- As **Velocity** goes up, the whiff rate increases

Optimization: (+, +)

3D Plots

Slider



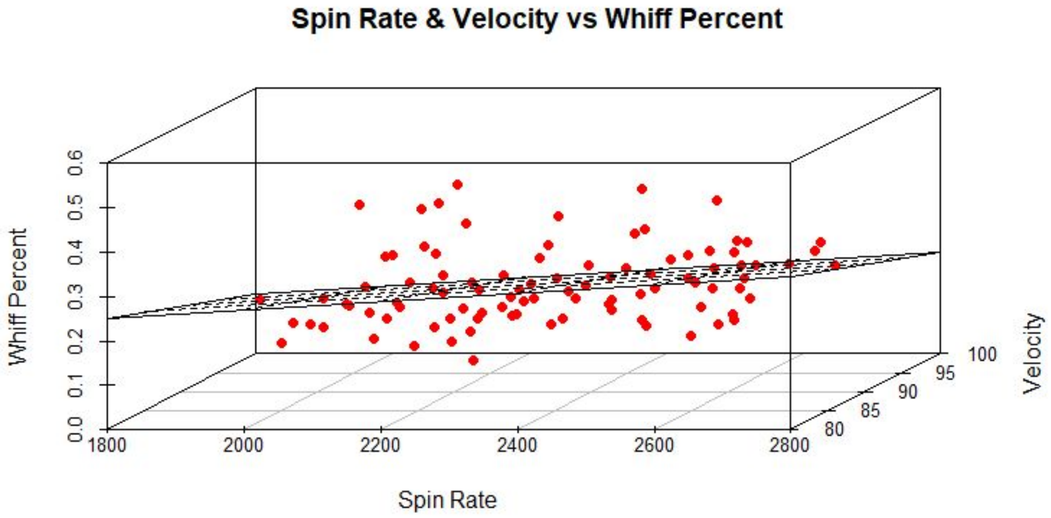
Trend:

- As **Spin Rate** goes up, the whiff rate increases
- As **Velocity** goes up, the whiff rate stays constant

Optimization: (+, null)

3D Plots

Cutter



Trend:

- As **Spin Rate** goes up, the whiff rate increases
- As **Velocity** goes up, the whiff rate decreases

Optimization: (+, -)

Mike Minor Z-scores (amongst MLB pitchers)

4-seamer:

- Spin rate: 2.26
- Velocity: -0.49

Curveball:

- Spin rate: -0.13
- Velocity: 0.54

Changeup:

- Spin rate: 1.93
- Velocity: 0.57

Slider:

- Spin rate: 1.28
- Velocity: 0.46

Practical Optimization: Minor

To maximize Minor's whiff rate against batters:

4 Seamer:

- Already high spin rate
- Increase his **velocity**

Curveball:

- Spin rate has minimal effect
- Increase his **velocity**

Changeup:

- Already high spin rate
- Increase his **velocity**

Slider:

- Increase his **spin rate** a bit
- Velocity has minimal effect

Nathan Eovaldi Z-scores (amongst MLB pitchers)

4-seamer:

- Spin rate: -0.73
- Velocity: 1.65

Curveball:

- Spin rate: -1.51
- Velocity: 0.48

Cutter:

- Spin rate: -0.05
- Velocity: 1.79

Slider:

- Spin rate: -0.84
- Velocity: -0.28

Practical Optimization: Eovaldi

To maximize Eovaldi's whiff rate against batters:

4 Seamer:

- Increase his **spin rate**
- Already high velocity

Curveball:

- Spin rate has minimal effect
- Increase his **velocity**

Cutter:

- Increase his **spin rate**
- Decrease his **velocity*****

Slider:

- Increase his **spin rate**
- Velocity has minimal effect

Conclusion

Our Conclusion: ***The best pitches have both high velocity and spin rate***

That's why Gerrit Cole and Justin Verlander are great while Mike Minor and Nathan Eovaldi are average pitchers

THANKS FOR LISTENING