UCLA College | Physical Sciences
Statistics & Data Science

520 Portola Plaza
8125 Math Sciences Building
Box 951554
Los Angeles, CA 90095

Balancing Performance and Longevity: Modeling Injury Risk in MLB Pitcher

Oscar Su, John Tan, Hannie Xie, Patrick Xue, Erica Yee (UCLA)

KEYWORDS — Baseball, Injuries, Performance, Pitches

1. Variables vs. Long Career Length

H_0: *all key indicators of pitchers whose career life is short (< 5 years) are the same as pitchers whose career life is long*

H_A: *there is at least one indicator that is different between the two types of pitchers*

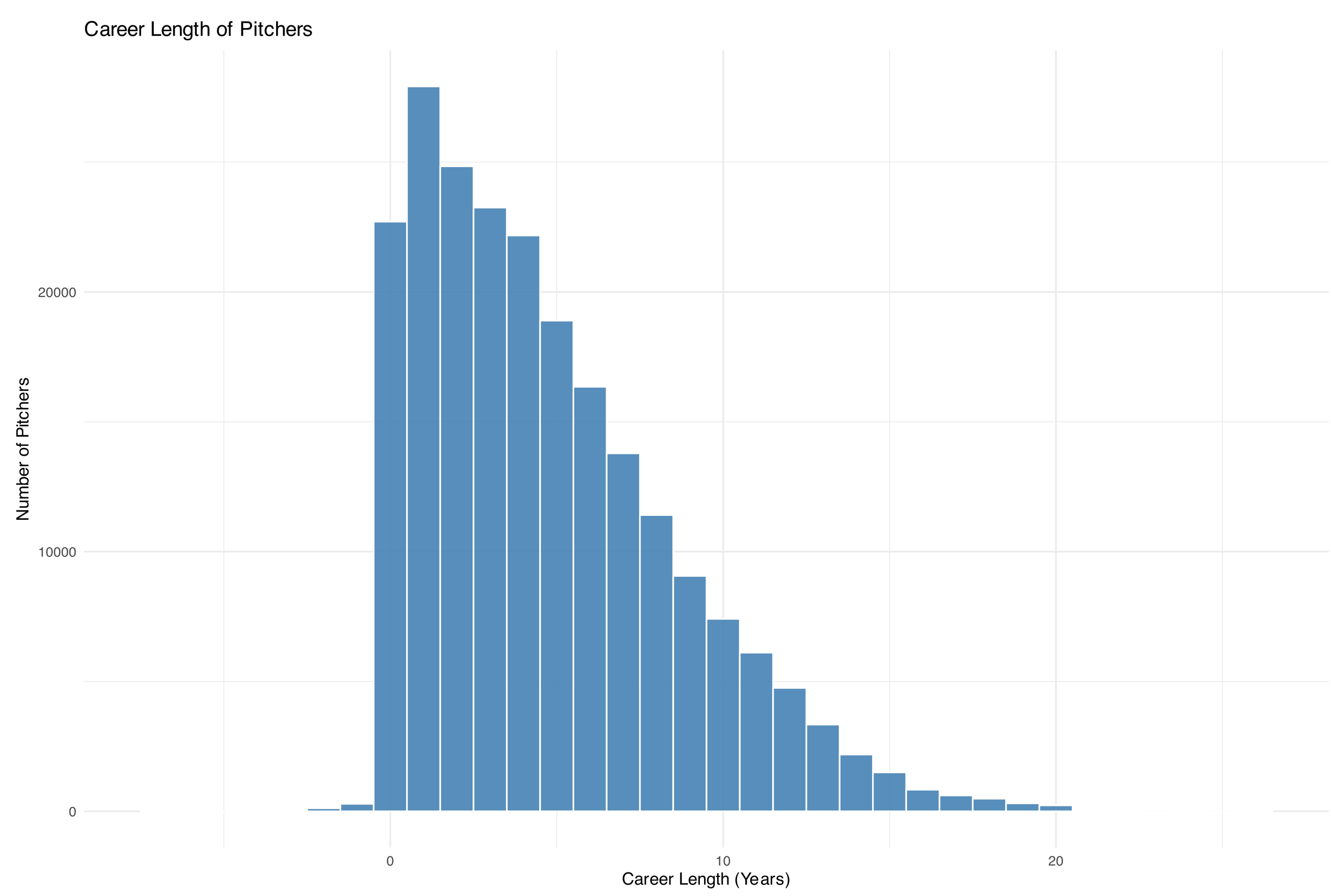


Table 1: t-test

Variable	P-Value	Mean (Short Career)	Mean (Long Career)
Height	0.58603	74.279	74.476
Weight	0.73358	213.162	214.397
Total Days Injured	0.56250	4.191	5.952
K_per_9	0.28747	7.001	7.566
BB_per_9	0.01740	4.785	3.624
HR_per_9	0.09470	1.381	0.990
Avg Fastball Overall	0.00095	90.983	92.426
Avg Breaking Overall	0.13390	80.489	81.409
Avg Offspeed Overall	0.01660	83.003	84.334
Avg Pitches Per Game	0.05437	33.760	42.903
Prop_Fastball	0.25529	0.570	0.590
Prop_Breaking	0.81394	0.205	0.201
Prop_Offspeed	0.63333	0.089	0.082

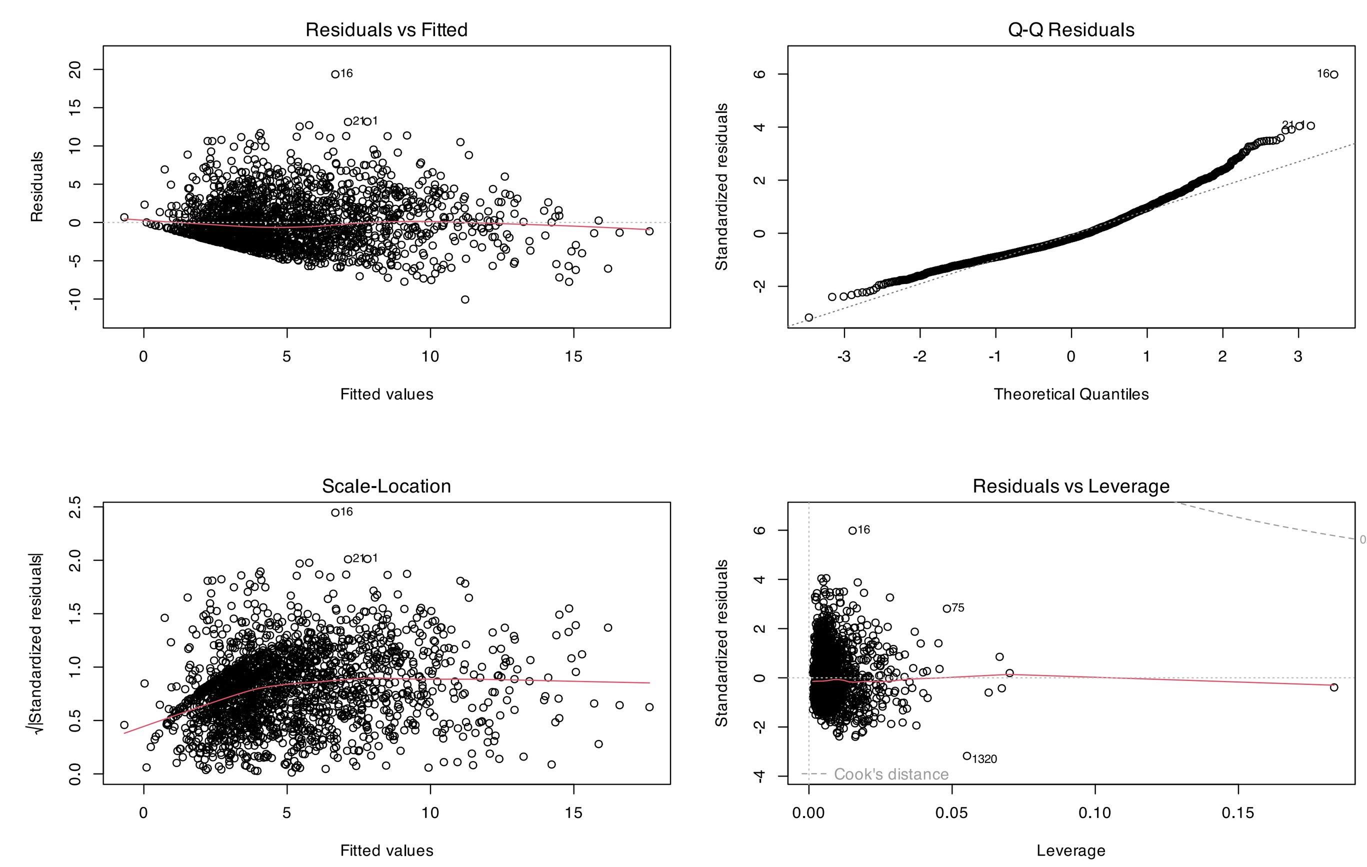
The t-tests compare pitchers with short versus long careers across multiple physical, performance, and pitch-mix variables. Significant p-values indicate meaningful differences between the two groups. Based on the updated results, most variables show statistically significant differences. In particular, pitchers with longer careers tend to throw more pitches per game, have higher average fastball and offspeed velocities, and walk fewer batters (lower BB_per_9). They also show differences in pitch mix, such as a lower proportion of fastballs.

From the table above, we see that BB_per_9, Avg_Fastball_Overall, Avg_Offspeed_Overall, and Prop_Breaking have p-values below 0.05, meaning these variables differ significantly between short- and long-career pitchers. In contrast, Height, Weight, Total_Days_Injured, K_per_9, HR_per_9, Avg_Breaking_Overall, Avg_Pitches_Per_Game, and Prop_Offspeed have p-values above 0.05, indicating no statistical evidence that these variables differ between the groups.

Then we do the stepwise model to achieve the best Final Model.

Table 2: Stepwise Process (backward), AIC

Step	Model Formula Terms Remaining	AIC
0	Full model (18 predictors)	261.98
1	Remove Avg_Offspeed_Overall	259.98
2	Remove Prop_Offspeed	257.99
3	Remove Height	256.00
4	Remove HR_per_9	254.07
5	Remove Avg_Breaking_Overall	252.29
6	Remove Total_K	250.59
7	Remove Total_Days_Injured	248.82
8	Remove Avg_Pitches_Per_Game	247.16
9	Remove Total_BB	245.36
10	Remove Total_HR	243.52
11	Remove Prop_Fastball	242.17
12	Remove Prop_Breaking	240.85
13	Remove Total_Pitches_Career	239.96
14	Remove K_per_9	238.84
15	Remove Weight	238.69
16	Final model	238.69



The final model is

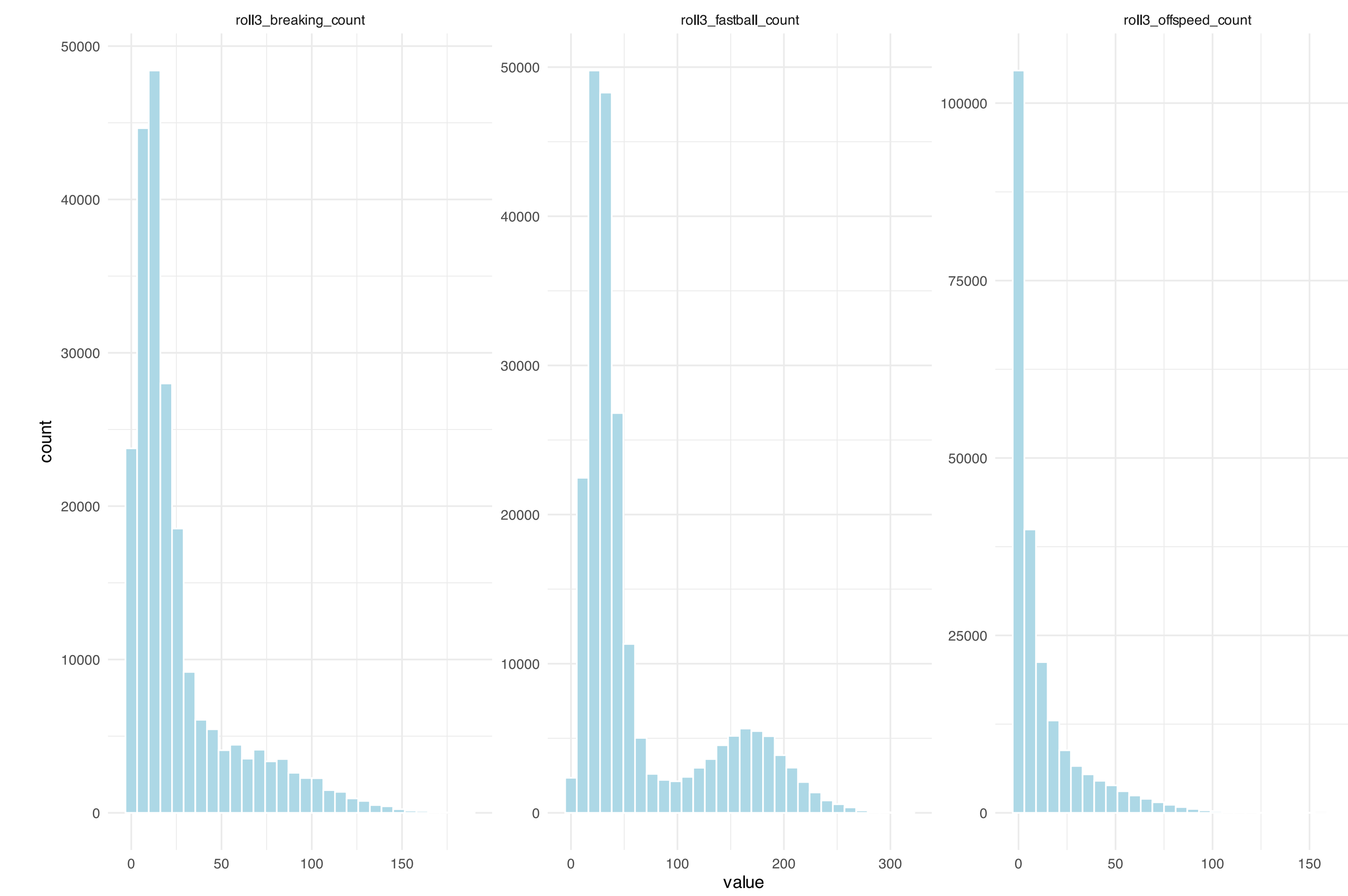
$$\text{Career Length} = -44.2 + 0.00684\,\text{Total Outs} - 0.244\,\text{BB\_per\_9} + 0.533\,\text{Avg Fastball Overall}$$

2. Different Types of Pitches vs. Injury & Performance

Motivation

- Pitch limits are widely used but not foolproof
- Maximum pitch counts per game have steadily dropped
- Routine workload hasn't actually decreased

Analysis



- *All three pitch-type categories are extremely right-skewed, with the vast majority of observations clustered near zero. This makes sense: most pitchers throw relatively few of any particular pitch type in any given 3-game rolling window.*

- *Long right tails are present for every pitch type, indicating that a smaller group of pitchers throw these pitches very frequently within short time spans.*

- *Fastballs have the highest counts and longest tails, consistent with being the most frequently thrown MLB pitches.*

= Logistic Regression Analysis

```
Call:
glm(formula = fml, family = "binomial", data = new_players_pitch_count)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.764673   0.535299 -10.769  < 2e-16 ***
roll3_fastball_count  0.029684   0.009787   3.033  0.00242 **
roll3_breaking_count -0.020956   0.023737  -0.883  0.37731
roll3_offspeed_count -0.031495   0.033183  -0.949  0.34255
rest_days             0.018842   0.018428   1.022  0.30655
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 265.19  on 2709  degrees of freedom
Residual deviance: 254.30  on 2705  degrees of freedom
AIC: 264.3

Number of Fisher Scoring iterations: 8
```

- *Throwing more fastballs increases injury risk* Your model suggests that each additional fastball thrown in the last 3 games raises the probability of injury by a small but statistically significant amount.

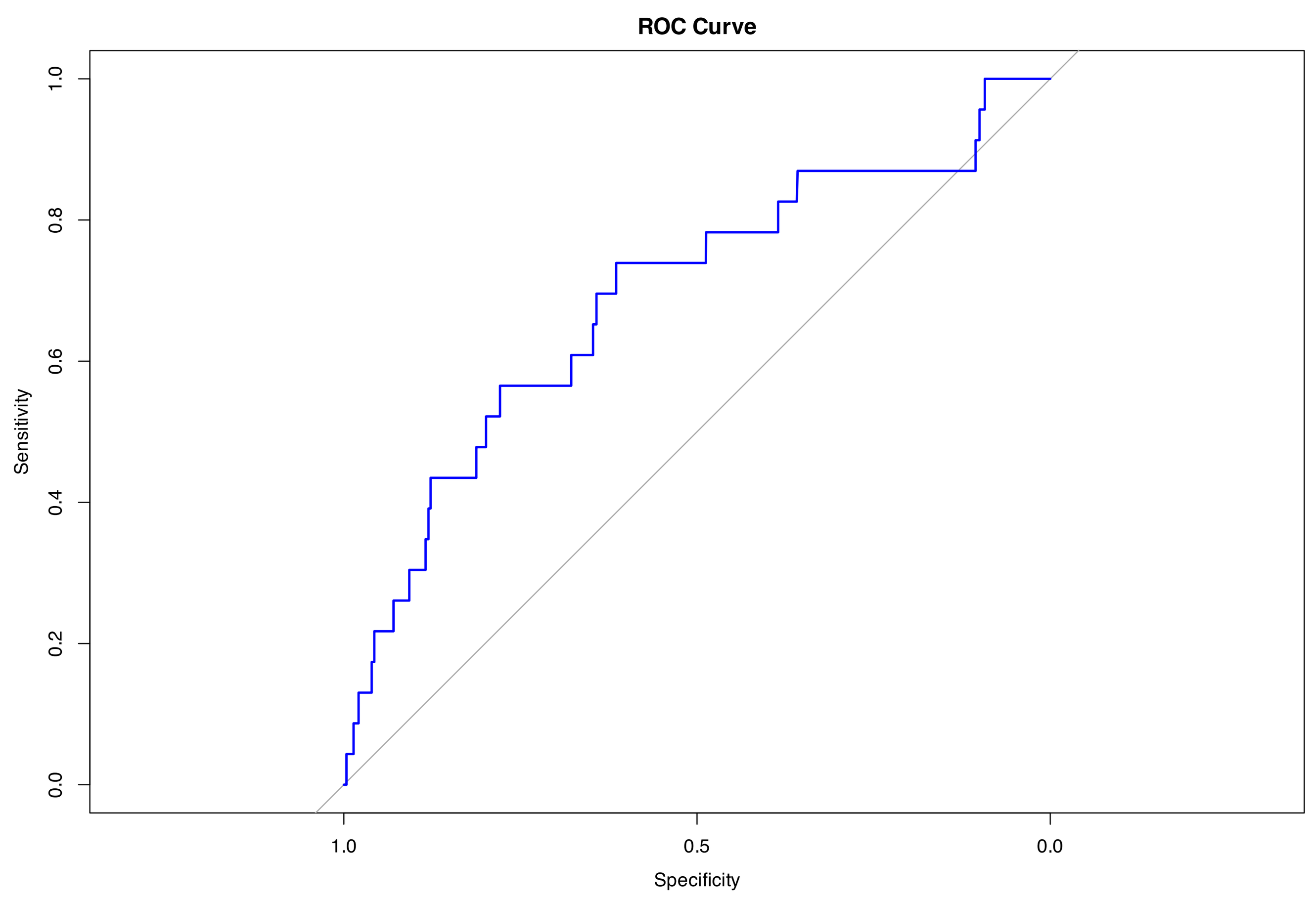
- *Breaking and offspeed pitches do not significantly affect injury risk* Their coefficients are small and non-significant, meaning there is no strong evidence that these pitch types contribute to injuries in this dataset.

- *Fastballs may increase injury risk because they impose higher mechanical stress* Fastballs generally require greater force production and higher arm speed, which can amplify elbow and shoulder loading.

- *Breaking and offspeed pitches may not matter for new players* These pitch types may generate less peak stress or be thrown less frequently, so their effects may not appear in a short-term injury window.

- *Rest days do not significantly reduce injury risk* The lack of statistical effect could mean that younger pitchers recover quickly, or that the rest-days measure is too coarse to capture true recovery patterns.*

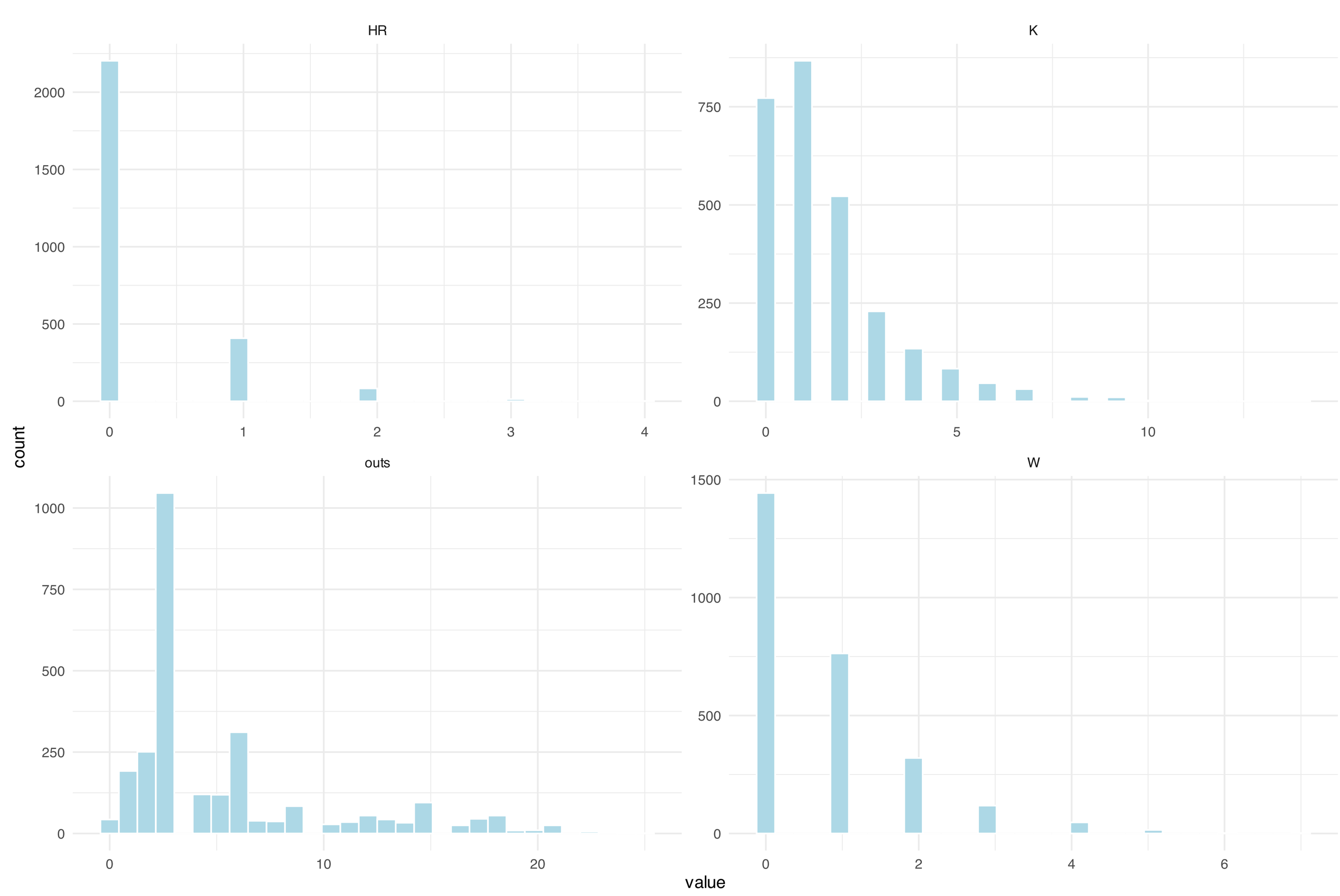
```
Area under the curve: 0.6895
```



Pitches Types vs. Performance

Hypothesis

- *H_0: all pitch types are not correlated with pitcher performance - H_A: there is at least one pitch type that is correlated to better pitcher performance*



Here we use Poisson regression model.

The model examines how pitch-type counts relate to the performance outcome

```
Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.4184550  0.0282053 -14.836  < 2e-16 ***
fastball_count  0.0188992  0.0009346  20.221  < 2e-16 ***
breaking_count  0.0284338  0.0017070  16.657  < 2e-16 ***
offspeed_count  0.0194577  0.0024030   8.097 5.61e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.1  on 2709  degrees of freedom
Residual deviance: 2765.9  on 2706  degrees of freedom
AIC: 7675.5

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.229033   0.041485 -29.626  < 2e-16 ***
fastball_count  0.025095   0.001324  18.955  < 2e-16 ***
breaking_count  0.021017   0.002574   8.164 3.23e-16 ***
offspeed_count  0.013711   0.003470   3.951 7.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3675.3  on 2709  degrees of freedom
Residual deviance: 2582.6  on 2706  degrees of freedom
AIC: 5526.6

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.538112   0.078153 -32.476  < 2e-16 ***
fastball_count  0.019012   0.002372   8.014 1.11e-15 ***
breaking_count  0.032527   0.004235   7.680 1.59e-14 ***
offspeed_count  0.038646   0.005669   6.817 9.26e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2177.3  on 2709  degrees of freedom
Residual deviance: 1757.7  on 2706  degrees of freedom
AIC: 2847.1

Number of Fisher Scoring iterations: 6
```

```
Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.7911495  0.0152265   51.96   <2e-16 ***
fastball_count 0.0217320  0.0004930   44.09   <2e-16 ***
breaking_count 0.0228794  0.0009356   24.45   <2e-16 ***
offspeed_count 0.0243477  0.0012328   19.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9543.7  on 2709  degrees of freedom
Residual deviance: 1693.1  on 2706  degrees of freedom
AIC: 10574

Number of Fisher Scoring iterations: 4
```

- *All pitch types significantly increase the expected number of events being modeled.* Every coefficient is positive and highly significant, meaning throwing more pitches of any type results in a higher expected count of the outcome.

- *Breaking pitches generally have the strongest effect, followed by fastballs, then offspeed pitches.* The coefficients for breaking pitches are consistently the largest across models, meaning a 1-unit increase in breaking pitches leads to the biggest increase in the event count.

Results

- *Pitchers improve performance when they throw more pitches, but doing so also raises injury risk, especially when relying heavily on breaking balls;*
- *This trade-off highlights the central tension between short-term performance gains and long-term health, emphasizing the importance of pitch-type management and load monitoring;*
- *The models agree with many research: breaking pitches are harder on the arm, and high-volume workloads will add up the injury risk.*