Uncovering the Metrics That Matter:

A Linear Modeling Approach to Hamstring Injury Risk in College Football

Jinwoo Choi
University of Arizona
October 12,2025

Abstract

This study aims to investigate the key physical and performance-based metrics that are collected from the advanced sports science technology that tracks athlete's performance. Using the metrics collected by this product, a linear regression model was developed to identify which variables most strongly predict injury occurrence .Predictor variables included measures such as total duration, player load, high-speed distance, acceleration and deceleration counts, total contacts and maximum velocity. The model was trained and validated to evaluate its predictive accuracy and to rank feature importance. Results highlight the metrics most associated with elevated injury risk, providing actionable insights for athletic training staff to optimize workload management and injury prevention strategies. This analysis demonstrates how data-driven modeling can support evidence-based decision-making in collegiate sports performance and health management. These results provide practical insights for sports scientists and coaches, emphasizing the importance of monitoring high-speed exposure and balancing workload intensity to optimize performance and reduce soft-tissue injury risk.

Introduction

At the competitive level of collegiate football, a program's success is ultimately measured by its ability to win games. Maintaining consistent access to top-performing athletes is therefore essential, underscoring the importance of minimizing injuries and keeping key players on the field. The physical demand of training and competition in a sport like Football can be overload for athletes, which most programs have started investing heavily in Sports Science departments in recent years to ensure the players are well-taken care of, optimize their performance, and most importantly to prevent injuries. As player-tracking technologies and performance analytics became more advanced, the sports science industry has been able to benefit from visualizing insights in workload, movement, and metrics. While certain injuries are unavoidable, such as those resulting from high-impact collisions, awkward landings, or underlying health conditions—many performance-related injuries, like hamstring strains, can be mitigated through proper training, conditioning, and workload management.

This project examines which training and workload metrics are most strongly associated with hamstring performance-related injuries in collegiate football athletes, using a linear regression approach. The analysis intentionally excludes external factors such as genetics, lifestyle, and diet, and focuses instead on indicators of training load within the program's structure that may contribute to physical strain and injury risk. The goal is to move beyond descriptive summaries toward actionable insights that enable coaching and sports science staff to better manage player workloads and support performance and longevity.

Dataset

The dataset was obtained from the University of Arizona's NCAA Division I football program, the Arizona Wildcat Football team, and was collected using the Catapult Vector 8[1] athlete monitoring system. Data were recorded during both practice sessions and competitive games, covering the period from the start of fall camp on July 31, 2025, through the October 4, 2025 game against Oklahoma State University, representing the early portion of the 2025–2026 NCAA football season. The dataset used for this analysis had been previously cleaned prior to modeling. The initial data set

[1] Catapult. (n.d.). *Vector 8*. Retrieved October 12, 2025, from https://www.catapult.com/vector8

has 2873 rows with 15 different columns, and a summary of each column values in the dataset is presented below.

- Player: Name of the individual athlete.
- Position_Name: Categorical value of players' position
- Velocity_Band_5_Total_Effort_Count: Number of instances in which the athlete reached between 80% and 95% of their maximum velocity during a session.
- Velocity_Band_5_Total_Distance: Total distance (in yards) covered by the athlete while moving at 80%–95% of their maximum velocity.
- Velocity_Band_6_Total_Effort_Count: Number of instances in which the athlete exceeded 95% of their maximum velocity during a session.
- Velocity_Band_6_Total_Distance: Total distance (in yards) covered by the athlete while exceeding 95% of their maximum velocity.
- Distance: Total distance (in yards) covered by the athlete during the session.
- Player_Load: A composite measure of total physical exertion, calculated by Catapult's inertial measurement algorithm.
- PL_Min: Player load normalized per minute of activity, representing intensity relative to session duration.
- Hi_Accel: Total count of high-intensity acceleration efforts.
- Hi_Decel: Total count of high-intensity deceleration efforts.
- Date: The date on which the practice or game session occurred.
- Game: Binary indicator denoting session type (1 = game, 0 = practice).
- High_Distance: Velocity_Band_5_Total_Distance + Velocity_Band_6_Total_Distance
- High_dist_ratio: High_Distance / (Distance - 500)

Following picture below shows the first 5 row of the data:

index	Player	Velocity_Band_5_Total_Effort_Count	Velocity_Band_5_Total_Distance	Velocity_Band_6_Total_Effort_Count	Velocity_Band_6_Total_Distance
0	Player_1	0	0.0	0	0.0
1	Player_2	0	0.0	0	0.0
2	Player_3	2	9.908	1	18.93
3	Player_4	1	8.005	0	0.0
4	Player_5	0	0.0	0	0.0

Distance	Player_Load	PL_Min	Hi_Accel	Hi_Decel	Date	Game	DB	WR	LB	RB	TE	DL	OL	High_Distance
2426.388	376.466	3.062	0	0	7/31/25	0	0	0	0	0	0	0	1	0.0
3138.285	297.796	2.953	0	0	7/31/25	0	0	0	0	0	0	0	0	0.0
2422.083	386.68	3.132	2	3	7/31/25	0	0	0	0	0	0	1	0	28.838
2761.8	429.922	3.482	1	2	7/31/25	0	0	0	0	0	0	1	0	8.005
2446.914	302.858	2.745	4	5	7/31/25	0	0	0	0	0	1	0	0	0.0

Following picture below shows the statistical summary of each column:

	Velocity_Band_5_Total_Effort_Count	Velocity_Band_5_Total_Distance	Velocity_Band_6_Total_Effort_Count	Velocity_Band_6_Total_Distance	Distance	Player_Load	PL_Min	Hi_Accel	Hi_Decel
count	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000
mean	0.780962	13.824022	0.040757	0.486045	3635.984379	362.627489	3.052476	4.119642	2.362346
std	1.433475	29.580384	0.226286	2.874398	1619.208647	163.313071	1.093815	5.547902	3.190872
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.123000	0.013000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	2508.438500	255.534500	2.375000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	3666.088000	368.580000	3.052000	2.000000	1.000000
75%	1.000000	18.432500	0.000000	0.000000	4815.075500	472.056000	3.771500	7.000000	4.000000
max	12.000000	346.085000	3.000000	52.122000	8661.729000	941.392000	6.579000	35.000000	24.000000

Game	DB	WR	LB	RB	TE	DL	OL	High_Distance	High_dist_ratio
3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000	3803.000000
0.095714	0.206679	0.155930	0.133579	0.067052	0.066001	0.146463	0.172495	14.310067	0.003834
0.294237	0.404976	0.362836	0.340244	0.250145	0.248316	0.353617	0.377860	30.553189	0.008740
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.008841
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	18.898000	0.004618
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	353.608000	0.124252

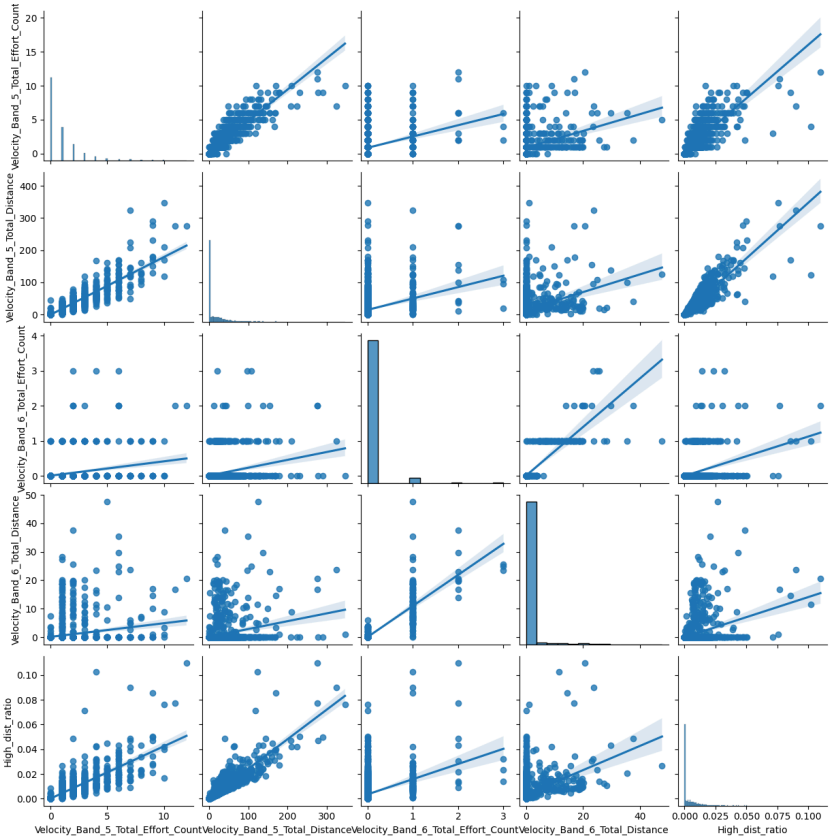
Pair Plots:

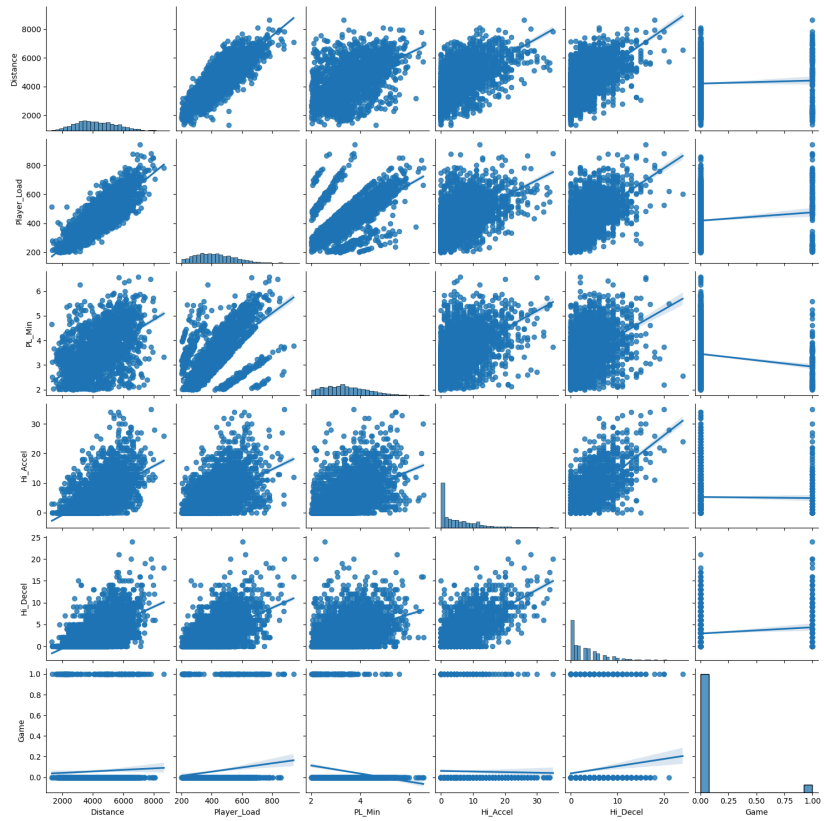
Now that we have the basic concept of the data, we will begin to look at the relationship between each variable by constructing pair plots. Since the data consists of many columns, I will split the pair plot into 2 parts, one being directly related to high-speed in sports context, such as:

- `Velocity_Band_5_Total_Effort_Count`
- `Velocity_Band_5_Total_Distance`
- `Velocity_Band_6_Total_Effort_Count`
- `Velocity_Band_6_Total_Distance`
- `High_dist_ratio`

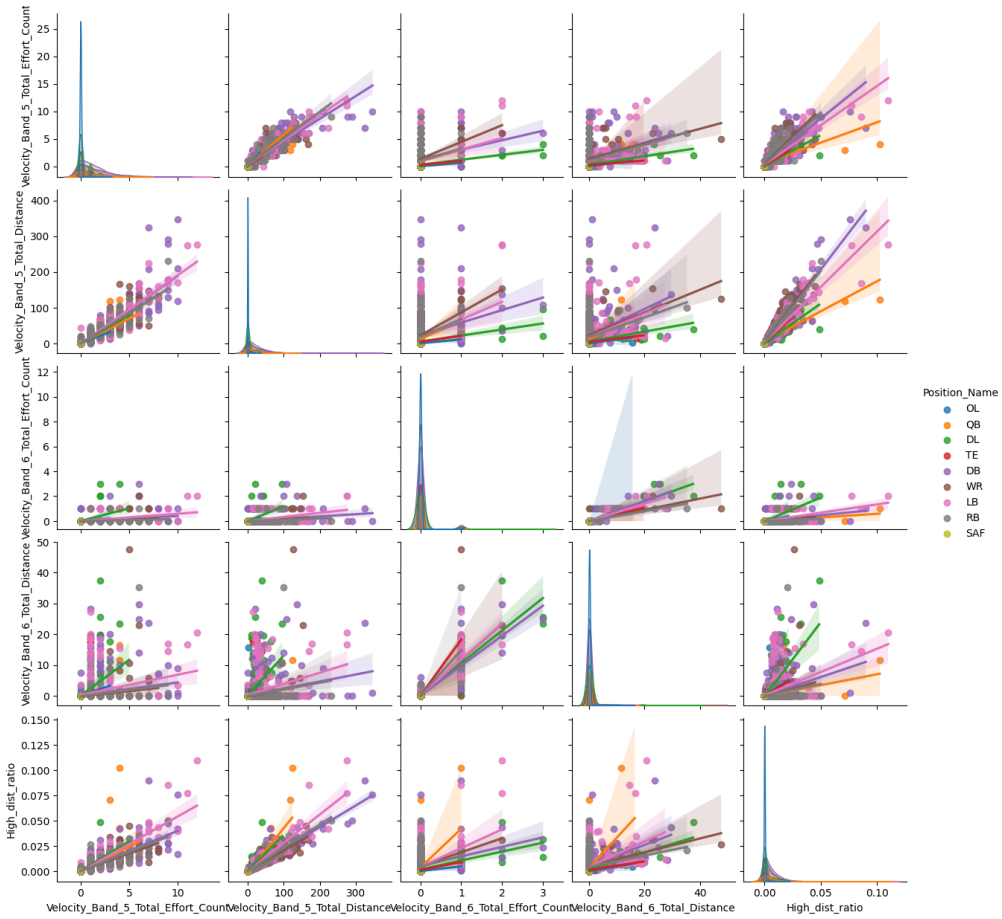
And the other pair plot consisting columns that could be externally related to high-speed under the sports context, such as:

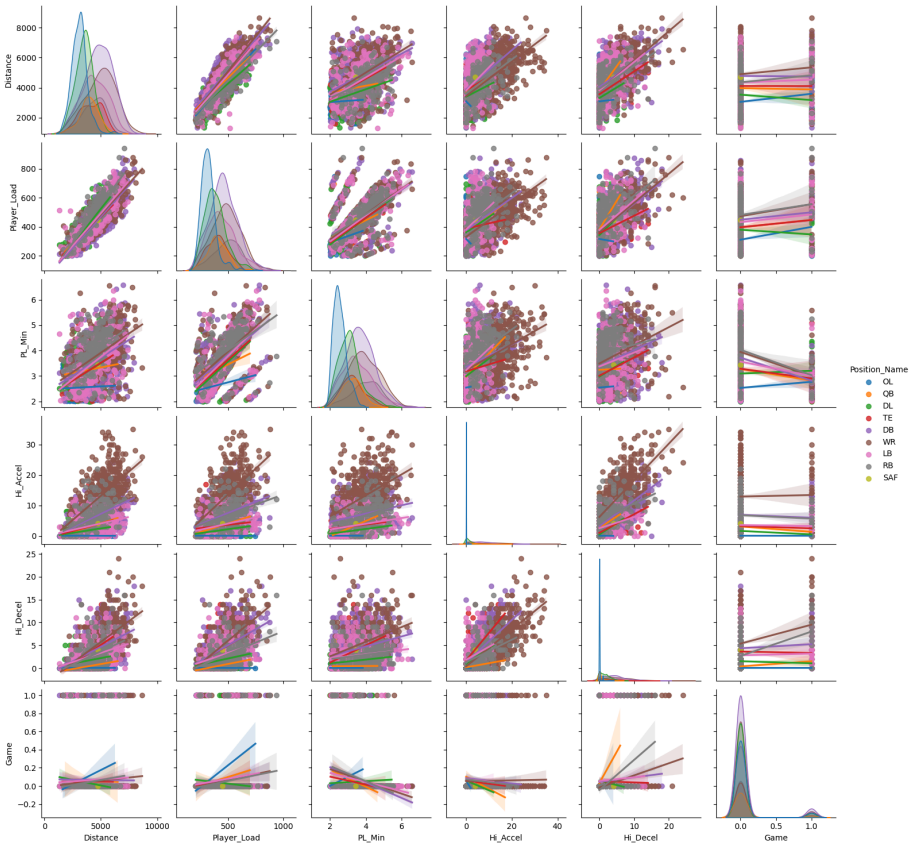
- `Distance`
- `Player_Load`
- `PL_Min`
- `Hi_Accel`
- `Hi_Decel`
- `Game`


Velocity_Band_5_Total_Effort_Count
Velocity_Band_5_Total_Distance
Velocity_Band_6_Total_Effort_Count
Velocity_Band_6_Total_Distance
High_dist_ratio


Distance
Player_Load
PL_Min
Hi_Accel
Hi_Decel
Game

Although this pair plot was able to give us some visualization, it is still hard to determine a clear relationship. Now I try to group by the position and plot different colors in each position to see if there could be a strong relationship within position groups.

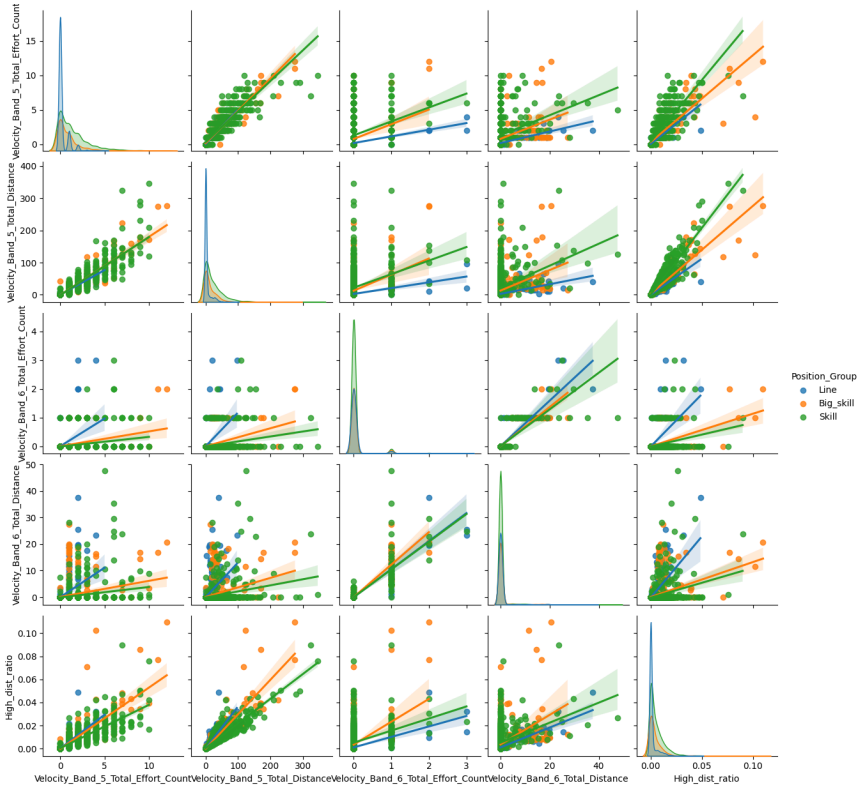


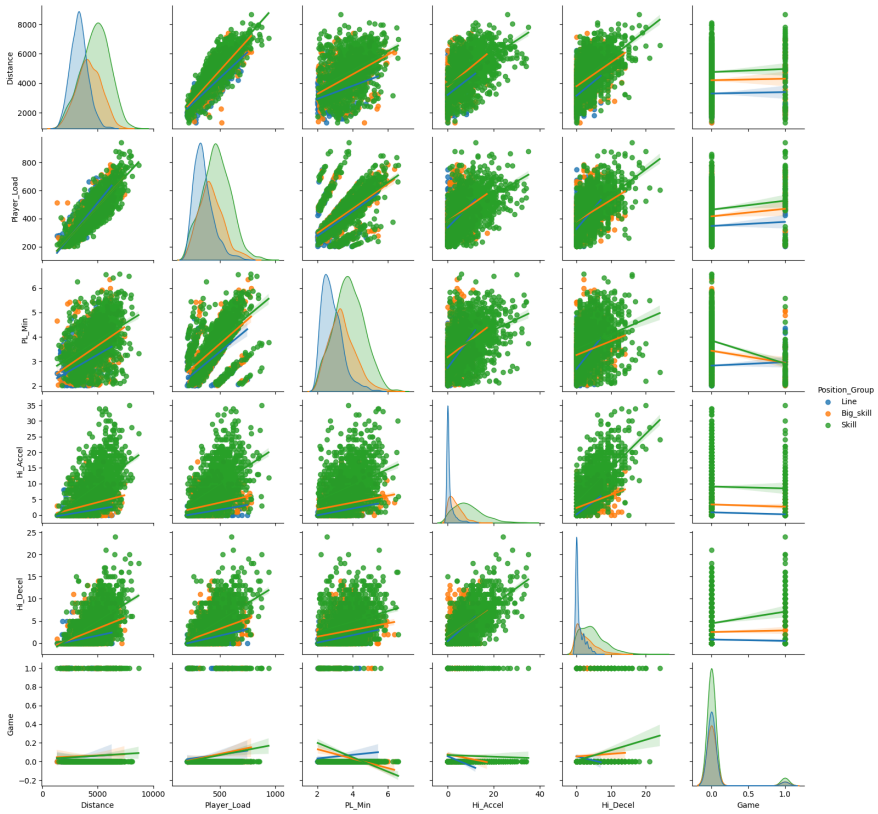


Although this graph has been more detailed and sees the trends within each position, the plot might have gotten more complicated to visualize. Finally, I will subset the dataset into 3 different data sets based on their position to see if creating groups based on their size will help us get a better understanding of the relationship within each category. Since Football is a very diverse sport with each position requiring different skillset, the positions are usually split into 3 different groups.

- Skill: Positions that prioritize high speed and explosiveness (WR, DB, and RB)
- Big-skill: Still requires speed but also bigger than the Skill group to absorb some physical contact (TE, LB, QB)
- Lineman: Requires pure physical contact and do not require high velocity (OL and DL)

Now that `Position_Group` has been created with values consisting of `['Skill', 'Big_Skill' 'Line']`, we will construct pair plots again.


Velocity_Band_5_Total_Effort_Count
Velocity_Band_5_Total_Distance
Velocity_Band_6_Total_Effort_Count
Velocity_Band_6_Total_Distance
High_dist_ratio
Position_Group
Line
Big_skill
Skill


Distance
Player_Load
PL_Min
Hi_Accel
Hi_Decel
Game
Position_Group
Line
Big_skill
Skill

We can now visualize the relationships between each feature and observe how the data varies across different position groups. The plot indicates that the Skill group generally occupies the upper quantiles of each feature, as expected, given that these positions involve frequent running and explosive movements. In contrast, the Lineman group tends to fall within the lower quantiles across all features, reflecting their reduced need for high-velocity actions. There seems to be potential influential points or outliers, but there are many points spread out, which makes it hard to determine, so further investigation might be needed. After examining the correlations among the features, we observe an overall positive trend with an approximately linear relationship. Although the degree of linearity requires further verification, the data appears well-suited for a linear modeling approach, which will be applied in the detailed methodology.

Methodology

To guide the analytical approach, I consulted with the Sports Science department of the Arizona Wildcats Football program to better understand which performance metrics are most relevant for identifying and mitigating hamstring injury. Through these discussions, staff emphasized that hamstring injuries are among the most common soft-tissue injuries observed in collegiate football and typically occur during high-speed sprinting activities. As a result, the high-speed distance metric was identified as a primary variable of interest for this study. Additionally, using the Catapult's product, I divided the high-speed distance into two levels based on intensity. The first group included movements between 80% and 95% of a player's maximum speed, and the second group included any movement at higher than 95% of maximum speed. This helped capture how much time players spent in different sprinting zones, which may relate to the chance of injury.

To develop the linear regression model for this project, the first step was to define the appropriate response variable. The goal was to measure the intensity of each practice session by examining the proportion of total distance covered at high speeds. To capture this, a high-speed ratio was created by combining the total distances from Velocity Band 5 (80–95% of max velocity) and Velocity Band 6 (above 95% of max velocity), then dividing that combined value by the total distance of the session, but subtract the total distance by 500 yd of constant value to exclude the non-intensive distance such as walking during practices, jogging to different stations during

practices, or walking in the sidelines, etc. This ratio represents the overall sprint workload for each player during a given session.

Throughout the visualization of the data, it's clear that different position groups have their own range of data values, which we will create a position group indicator value , the following columns have been created for this group indicator.

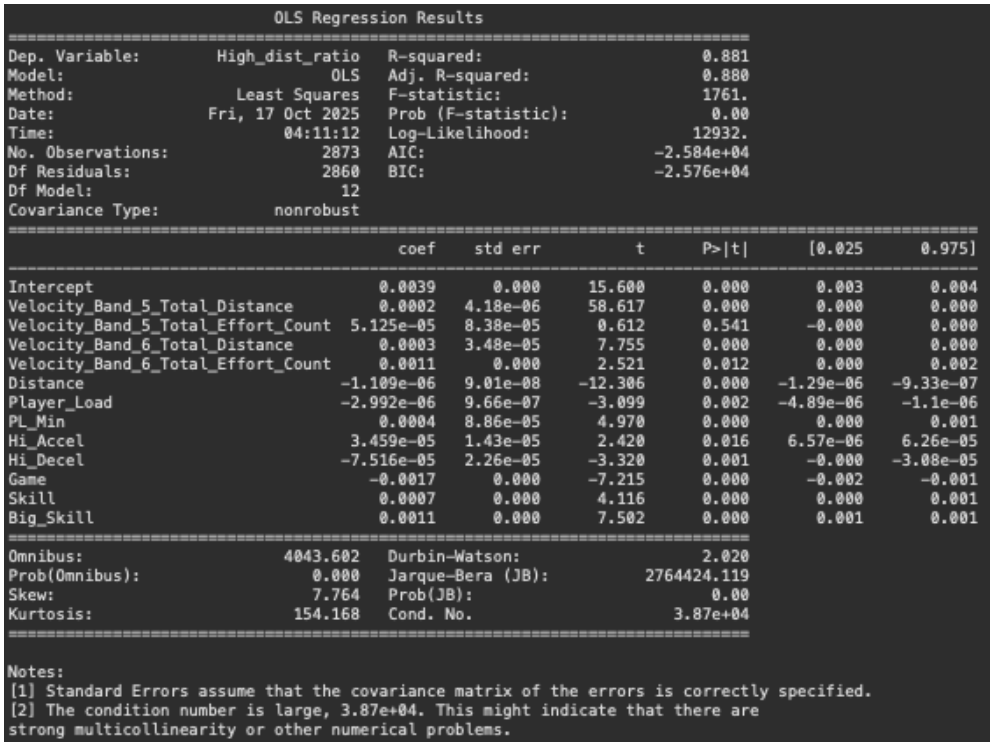
- Skill: Binary indicator denoting skill group (1 = Skill, 0 = Not Skill).
- Big_Skill: Binary indicator denoting skill group (1 = Big_Skill, 0 = Not Big_Skill).
- If both values are 0, the player is in the Line group.

Now we will begin creating the linear model, which we will begin with putting all the columns besides categorical values such as date, player's name, and player's position. The equation of the model is shown below.

$$\widehat{\text{High.dist.ratio}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{12} X_{12} + \epsilon$$

- X_1: Velocity_Band_5_Total_Distance
- X_2: Velocity_Band_5_Total_Effort_Count
- X_3: Velocity_Band_6_Total_Distance
- X_4: Velocity_Band_6_Total_Effort_Count
- X_5: Distance
- X_6: Player_Load
- X_7: PL_Min
- X_8: Hi_Accel
- X_9: Hi_Decel
- X_10: Game
- X_11: Skill
- X_12: Big_Skill

Here is the following model summary of this model:



OLS Regression Results

Dep. Variable:	High_dist_ratio	R-squared:	0.881
Model:	OLS	Adj. R-squared:	0.880
Method:	Least Squares	F-statistic:	1761.
Date:	Fri, 17 Oct 2025	Prob (F-statistic):	0.00
Time:	04:11:12	Log-Likelihood:	12932.
No. Observations:	2873	AIC:	-2.584e+04
Df Residuals:	2860	BIC:	-2.576e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0039	0.000	15.600	0.000	0.003	0.004
Velocity_Band_5_Total_Distance	0.0002	4.18e-06	58.617	0.000	0.000	0.000
Velocity_Band_5_Total_Effort_Count	5.125e-05	8.38e-05	0.612	0.541	-0.000	0.000
Velocity_Band_6_Total_Distance	0.0003	3.48e-05	7.755	0.000	0.000	0.000
Velocity_Band_6_Total_Effort_Count	0.0011	0.000	2.521	0.012	0.000	0.002
Distance	-1.109e-06	9.01e-08	-12.306	0.000	-1.29e-06	-9.33e-07
Player_Load	-2.992e-06	9.66e-07	-3.099	0.002	-4.89e-06	-1.1e-06
PL_Min	0.0004	8.86e-05	4.970	0.000	0.000	0.001
Hi_Accel	3.459e-05	1.43e-05	2.420	0.016	6.57e-06	6.26e-05
Hi_Decel	-7.516e-05	2.26e-05	-3.320	0.001	-0.000	-3.08e-05
Game	-0.0017	0.000	-7.215	0.000	-0.002	-0.001
Skill	0.0007	0.000	4.116	0.000	0.000	0.001
Big_Skill	0.0011	0.000	7.502	0.000	0.001	0.001

Omnibus:	4043.602	Durbin-Watson:	2.020
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2764424.119
Skew:	7.764	Prob(JB):	0.00
Kurtosis:	154.168	Cond. No.	3.87e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.87e+04. This might indicate that there are strong multicollinearity or other numerical problems.

The linear model demonstrated a strong fit, with an R-squared value of 0.881, indicating that approximately 88.1% of the variation in the high-speed distance ratio is explained by the selected predictors. While most variables showed statistically significant relationships, scoring significantly low p-values, the Velocity Band 5 and 6 Total Effort Count variables were surprisingly among the least significant predictors. This goes against the basic expectation that more high-speed efforts would lead to a higher proportion of high-speed running. It's possible that these variables overlap with other workload measures, such as total high-speed distance, which may already capture most of their effect. Conversely, variables such as total distance, player load, displayed negative relationships with the high-speed ratio, while the player load per minute had positive relationships. From a sports science perspective, this can be explained by fatigue and workload effects, players who accumulate greater total distance or load are less likely to sustain high-intensity sprinting, reflecting decreased sprint output over time, while the work density like player load per minute could affect the high-speed ratio proportionally. Similarly, higher deceleration counts suggest frequent stop-and-go movements, limiting opportunities to reach maximum velocity and therefore reducing the proportion of high-speed distance.

For the categorical variables, including Game and the position groups, the model produced significantly low p-values, indicating that these categories differ meaningfully from the reference group after controlling for all other predictors. The Game variable's low p-value suggests a statistically significant difference in high-speed ratio between game sessions and practices. Similarly, the position indicators show significant differences in high-speed ratio across player groups, implying that positional characteristics such as player size, role, and movement demands are relevant factors influencing high-speed performance.

After analyzing the relationship between each predictor and the response variable, it was determined that removing certain features could improve the model's overall performance. To begin the refinement process, the two least significant predictors, Velocity Band 5 Total Effort Count and Velocity Band 6 Total Effort Count were excluded from the model. The updated regression model, reflecting these adjustments, is summarized below.

OLS Regression Results

Dep. Variable:	High_dist_ratio	R-squared:	0.881
Model:	OLS	Adj. R-squared:	0.880
Method:	Least Squares	F-statistic:	2109.
Date:	Fri, 17 Oct 2025	Prob (F-statistic):	0.00
Time:	05:22:02	Log-Likelihood:	12929.
No. Observations:	2873	AIC:	−2.584e+04
Df Residuals:	2862	BIC:	−2.577e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0039	0.000	15.556	0.000	0.003	0.004
Velocity_Band_5_Total_Distance	0.0002	1.95e−06	126.960	0.000	0.000	0.000
Velocity_Band_6_Total_Distance	0.0003	1.78e−05	19.393	0.000	0.000	0.000
Distance	−1.099e−06	9e−08	−12.209	0.000	−1.28e−06	−9.22e−07
Player_Load	−3.157e−06	9.64e−07	−3.273	0.001	−5.05e−06	−1.27e−06
PL_Min	0.0005	8.85e−05	5.166	0.000	0.000	0.001
Hi_Accel	3.553e−05	1.42e−05	2.509	0.012	7.77e−06	6.33e−05
Hi_Decel	−7.39e−05	2.27e−05	−3.263	0.001	−0.000	−2.95e−05
Game	−0.0017	0.000	−7.285	0.000	−0.002	−0.001
Skill	0.0006	0.000	3.968	0.000	0.000	0.001
Big_Skill	0.0011	0.000	7.389	0.000	0.001	0.001

Omnibus:	4033.340	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2760342.735
Skew:	7.722	Prob(JB):	0.00
Kurtosis:	154.064	Cond. No.	2.37e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.37e+04. This might indicate that there are strong multicollinearity or other numerical problems.

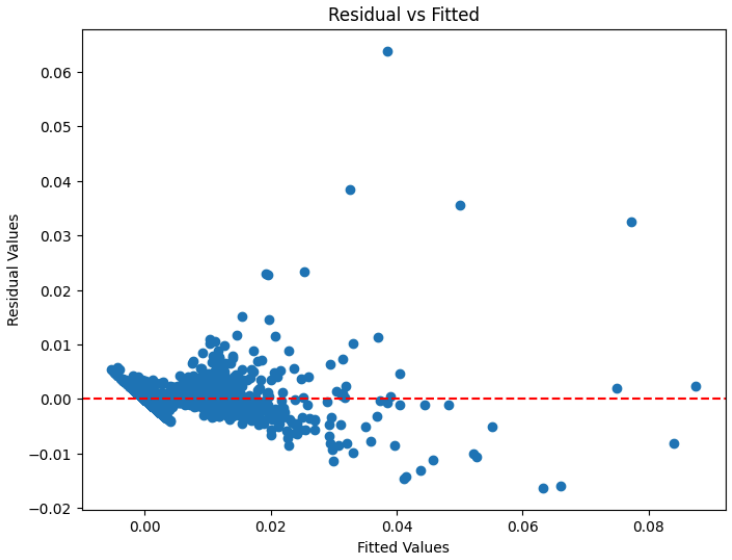
After removing the two least significant predictors (Velocity Band 5 and 6 Total Effort Count), the refined linear model continued to perform strongly with the same R-squared value. The results reinforce that high-speed distance metrics remain the most important drivers of sprint workload. Specifically, greater distances covered in Velocity Band 5 and Velocity Band 6 are strongly associated

with a higher proportion of high-speed running, which aligns with the expected physical factor linked between sprint exposure and session intensity.

Similarly with the previous model, total distance and player load both showed negative relationships with high-speed ratio, suggesting that as the overall workload and movement volume increase, athletes are less likely to maintain top-end sprinting intensity. This pattern reflects the practical reality that longer or more physically demanding sessions tend to cause fatigue accumulation, leading to reduced sprint efficiency. PL/Min and high acceleration counts were positively associated with high-speed ratio, indicating that sessions with higher intensity and explosive effort tend to produce greater high-speed running activity. Conversely, high deceleration counts were negatively related, implying that sessions involving frequent stop-and-go movements limit players' ability to reach and sustain higher sprint velocities. Overall, the new model suggests a better fit for this dataset compared to the previous model.

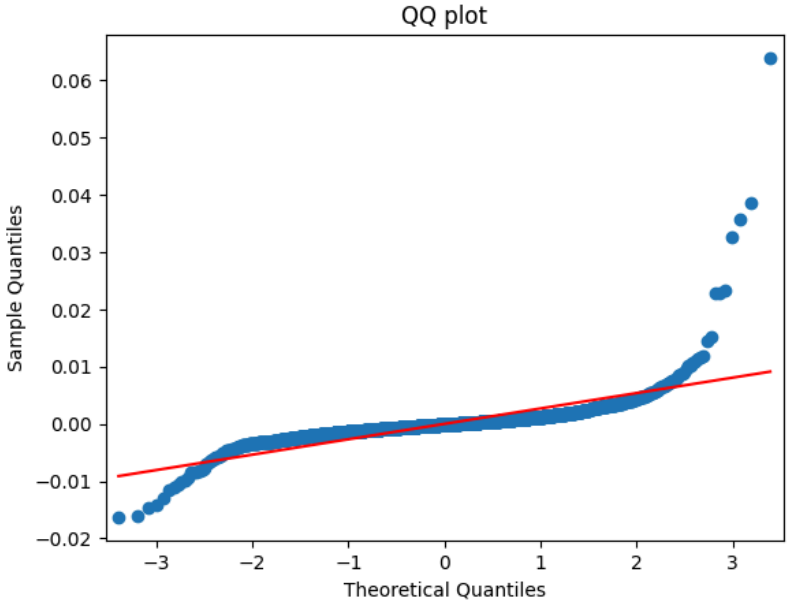
Diagnostics:

Residual vs Fitted & QQ Plot:



The residuals versus fitted values plot shows that most residuals are centered at the beginning, suggesting that the model captures the general trend of the data. However, a visible increase in residual spread at higher fitted values indicates mild heteroscedasticity. This means the model's prediction error tends to increase for sessions with higher high-speed ratios. Such behavior

is expected in performance datasets where variability naturally increases with workload intensity. Overall, the model does not meet basic linearity assumptions, but there could be possible transformation techniques to improve the non-constant variance issue.



The QQ plot of the residuals indicates that most data points align closely with the reference line, suggesting that residuals are approximately normally distributed in the central portion of the data. However, deviations at both tails, particularly in the upper tail, indicate the presence of a few extreme residuals, likely corresponding to sessions with unusually high high-speed ratios. This suggests mild right-skewness in the residual distribution. While this represents a minor concern from normality, it is not uncommon in sports performance data and does not substantially affect the reliability of the regression model given the sample size.

Additionally, calculating the normality through the Shapiro-Wilk Test Statistic, The test produced a statistic of 0.588 and a p-value of 0.000, indicating a significant deviation from normality. Since the p-value is below the conventional threshold of 0.05, the null hypothesis of normally distributed residuals is rejected. This suggests that the model's residuals are not perfectly normal, potentially due to the presence of outliers or skewed data. However, given the large sample size and the model's strong overall fit, this violation is unlikely to substantially affect the reliability of the regression estimates, as linear regression is generally robust to mild deviations from normality.

Outliers:

It is clear from the diagnostic plots that several outliers exist in the dataset, which may be affecting the accuracy of the model's predictions. To address this, I will begin by identifying these outliers using the Z-score method, which helps find data points that are far from the average value. Detecting these extreme values will make it easier to understand how they influence the model and whether they should be adjusted or removed for better performance.

```
Number of outliers: 38
Outlier indices: [ 130  476  673  701  777  820 1168 1178 1338 1427 1444 1451 1534 1760
 1762 1763 1823 1986 2008 2015 2020 2141 2178 2355 2406 2481 2588 2593
 2597 2644 2678 2785 2807 2820 2825 2826 2829 2853]
```

Results

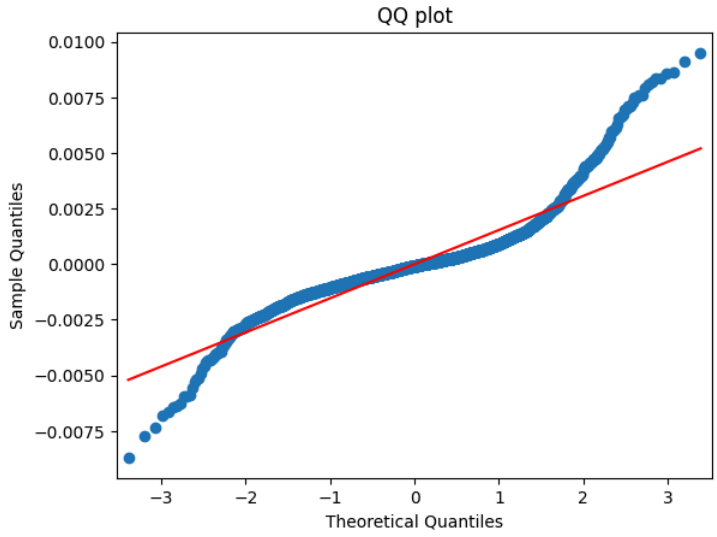
A total of 38 data points were identified as outliers that deviated substantially from the average values. After detecting these observations, they were removed from the dataset to reduce their influence on the model's performance. The updated regression model summary and the QQ plot after outlier removal is presented below.

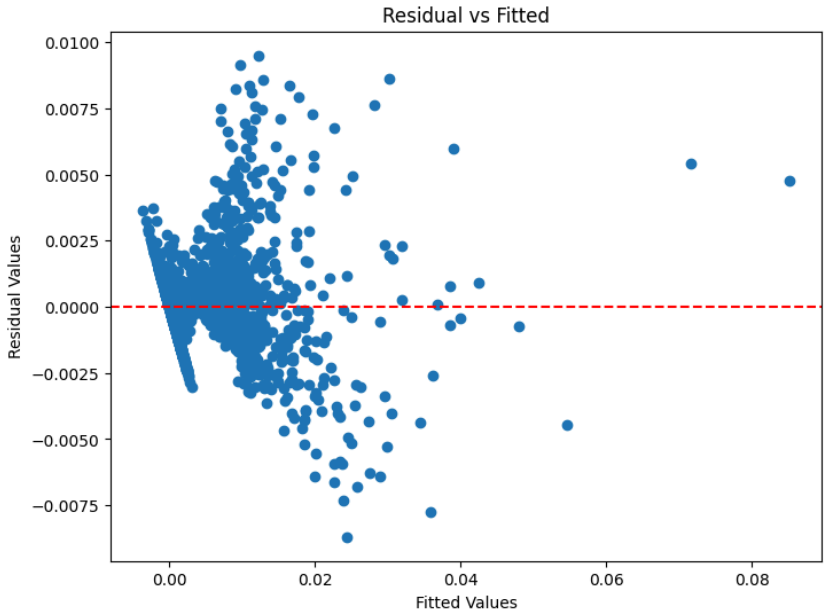
OLS Regression Results

Dep. Variable:	High_dist_ratio	R-squared:	0.939
Model:	OLS	Adj. R-squared:	0.939
Method:	Least Squares	F-statistic:	4348.
Date:	Fri, 17 Oct 2025	Prob (F-statistic):	0.00
Time:	06:06:32	Log-Likelihood:	14346.
No. Observations:	2835	AIC:	−2.867e+04
Df Residuals:	2824	BIC:	−2.860e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0030	0.000	20.911	0.000	0.003	0.003
Velocity_Band_5_Total_Distance	0.0002	1.31e−06	183.661	0.000	0.000	0.000
Velocity_Band_6_Total_Distance	0.0003	1.19e−05	26.738	0.000	0.000	0.000
Distance	−8.776e−07	5.2e−08	−16.878	0.000	−9.8e−07	−7.76e−07
Player_Load	−2.57e−07	5.58e−07	−0.460	0.645	−1.35e−06	8.38e−07
PL_Min	0.0001	5.14e−05	2.476	0.013	2.65e−05	0.000
Hi_Accel	2.017e−06	8.19e−06	0.246	0.806	−1.41e−05	1.81e−05
Hi_Decel	−3.042e−05	1.31e−05	−2.324	0.020	−5.61e−05	−4.75e−06
Game	−0.0014	0.000	−10.604	0.000	−0.002	−0.001
Skill	0.0007	9.35e−05	7.487	0.000	0.001	0.001
Big_Skill	0.0007	8.48e−05	8.681	0.000	0.001	0.001

Omnibus:	812.257	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6979.351
Skew:	1.109	Prob(JB):	0.00
Kurtosis:	10.360	Cond. No.	2.39e+04

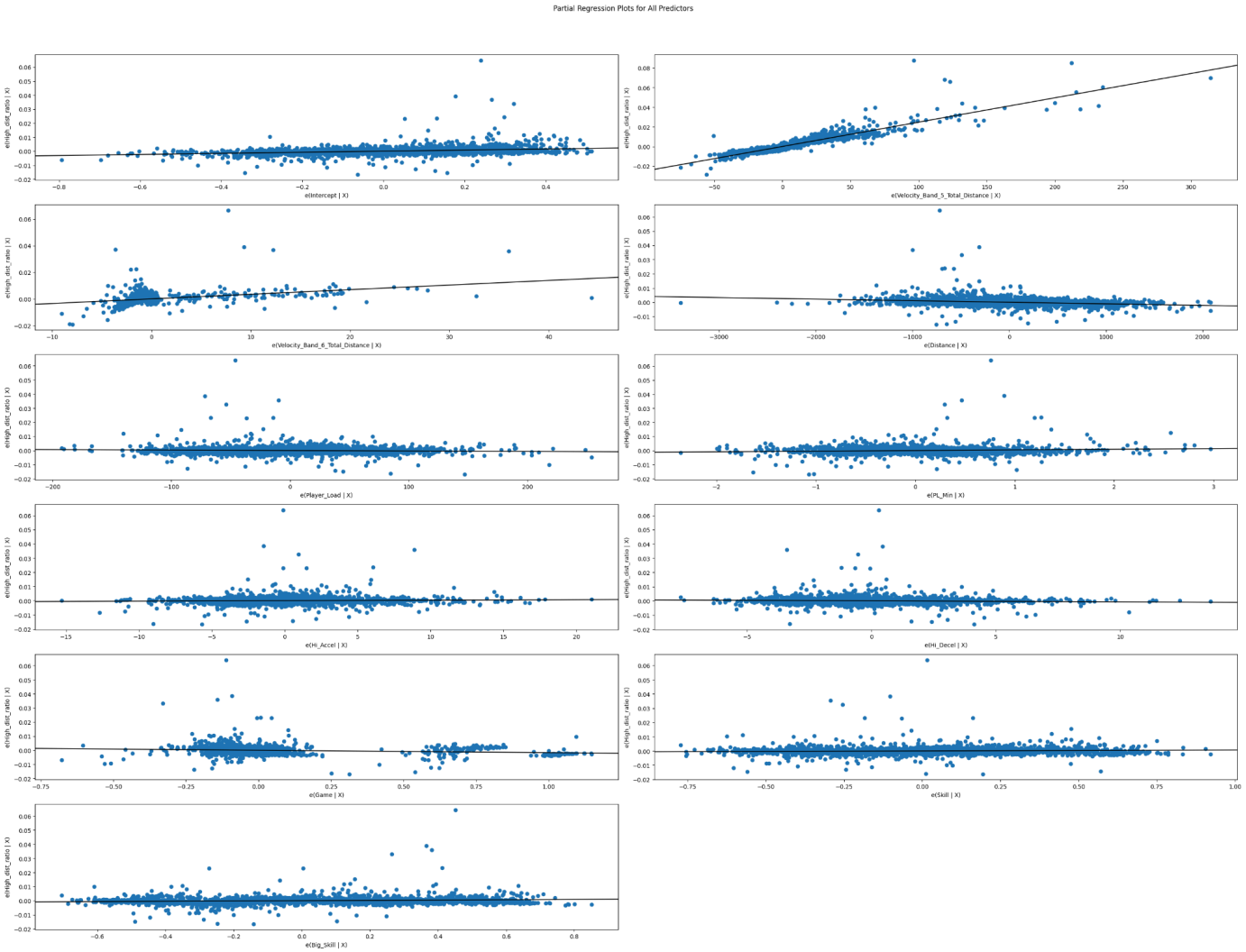




After removing the 38 identified outliers, the refined regression model showed a substantial improvement in performance, achieving an R-squared value of 0.939. The overall model remained highly significant with low F-statistic and a significant p-value, suggesting that removing extreme observations enhanced its reliability, along with the predictor variables remaining low p-values. Although the residual plot still showed spread out variance, the new QQ plot shows that the residuals now follow a much more linear pattern along the red reference line compared to the previous version. This indicates that the residuals are more normally distributed, meaning the model's assumptions of normality are now much better satisfied. The slight deviations at both ends

of the plot (suggest the presence of a few mild outliers, but these are not extreme enough to distort the model. Overall, this improved QQ plot suggests that removing or adjusting outliers helped stabilize the residual distribution and made the regression model more reliable.

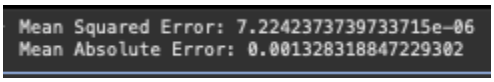
Partial Regression Plot:



The partial regression plots collectively show that variables related to high-speed distance have generally positive relationships with the high-speed ratio. Among these, Velocity Band 5 and 6 Total Distance displays the steepest slope, and the Total distance showing slight negative relationship, indicating a stronger influence on high-speed running intensity, while the fatigue from overload movement, suggesting weaker effects. Although some predictors exhibit only modest trends due to the small range of response values, the overall positive direction across multiple variables suggests that greater workload metrics are associated with higher proportions of

high-speed running. The position group indicator plots show the relationship between the high-speed distance ratio for each position group. These findings are consistent with on-field movement patterns and confirm that both workload and players' size are key factors influencing sprint intensity among collegiate football athletes.

Evaluating metrics:



The performance of this linear regression model was evaluated using the Mean Squared Error(MSE) and Mean Absolute Error(MAE) metric, which measures the average squared difference between the predicted and actual value of the high-speed distance ratio, and the average absolute difference between the predicted and actual value of the high-speed distance ratio. This model achieves the MSE value of 7.22 EE-5, and the MAE value of 1.33EE3 indicating that the errors are relatively small and that the model's fitted values align closely with the actual observed data. Through this metric, the low error value suggests that the model is highly effective and accurate at capturing the relationship between the workload and position features with the high-speed ratio. Overall, the MSE and MAE supports that dataset has a strong linear relationship that aligns with the linear regression model.

Conclusion

This project explored the relationships between various workload metrics and high-speed running performance in collegiate football athletes using a linear regression model. Through this model, the study identified which predictors had the strongest influence on high-speed performance, as well as those with comparatively weaker effects. Although the model demonstrated a strong fit and low prediction error, a couple of linear regression assumptions were not fully satisfied, suggesting that a more advanced modeling approach could better capture the complexity of the dataset. These findings provide valuable insights for sports scientists and coaching staff, emphasizing the importance of monitoring high-speed exposure, managing total workload, and balancing intensity with recovery to enhance performance while reducing the risk of soft-tissue injuries such as hamstring strains. In future work, I plan to extend this analysis by incorporating historical injury data and developing more sophisticated predictive models to estimate injury risk

more directly. This would enable a deeper investigation into the relationship between workload metrics and specific injuries, such as hamstring strains, at the individual athlete level.

Reference

- University of Arizona, Arizona Wildcat Football Team, Sports Science Department, Strength & Conditioning Program. (2025). *Player workload dataset collected using Catapult Vector 8* [Unpublished raw data]. Provided by Catapult Specialist & Assistant Director of Football Strength & Conditioning James Perez.

Code

- https://colab.research.google.com/drive/1VC1wBTaj2Sw7BZaomteZ3g-sEAS6sdAp?usp=sharing
- https://github.com/jinwoo1015/hamstring_injury_metrics