I. Introduction

In Major League Baseball (MLB), success extends beyond the field, reaching boardrooms where financial strategies and team productivity converge. This paper explores the pervasive impact of player injuries on MLB organizations' operations, extending from dugouts to financial cores. The persistent toll of players missing games is a formidable obstacle, hindering team productivity and eroding financial stability. MLB organizations must fortify against player injuries. Business Intelligence (BI) is key to informed decision-making and long-term sustainability. This BI project utilizes injury data, player age, and financial metrics to predict and prevent injuries, mitigating financial risks while enhancing team productivity. Figure 1 illustrates that MLB teams lose over $25 million per year due to player injuries, solely based on contract value and not including residual effects (ticket sales, marketing, fan engagement, etc.). This visually underscores the urgency of predictive analytics in player management. As the exploration unfolds, the goal is clear: empower MLB organizations with the foresight to navigate the unpredictable terrain of player injuries, ensuring sustained team productivity and financial resilience in the MLB's competitive arena.

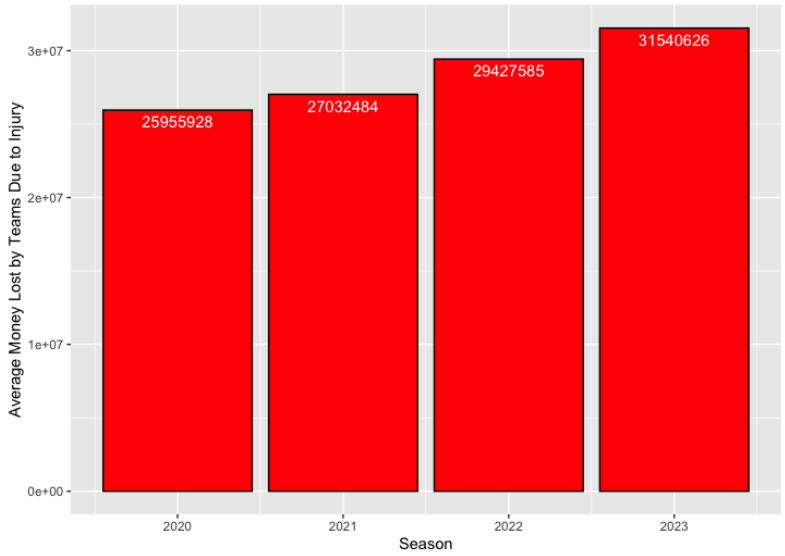


Figure 1: Bar Chart depicting Average Monetary Impact of Player Injuries on MLB Teams (Yearly)

II. Business Intelligence Methods

Guided by a comprehensive BI roadmap (Appendix 1), our project spans the 4-month MLB offseason. Led by a collaborative team, including myself, business sponsors, and the coaching staff, our primary goal is to deliver two key insights: Injury Analysis and Prediction, and Relationship Analysis. These insights are crucial for enhancing strategic decision-making processes within the organization. The project's success depends on active involvement from Business Sponsors, including the C-Suite, VP of Baseball Operations, and VP of Player Development. Their commitment not only steers the project but is crucial for realizing benefits. Business Constituencies, including team scouts, coaching staff, and the financial division, will optimize their responsibilities based on these insights. The project's impact extends to critical business processes, refining coaching strategies, free agency decisions, and strength and conditioning programs. Internally, business intelligence insights integrate into vital applications such as scouting reports and contract negotiations.

To gain useful insights, we focus on specific data subjects such as team payroll, player demographics, statistics, and team performance metrics. We gather raw data from various source systems such as FloQast (accounting software used by many MLB teams), wearable technology (yet to be implemented), and Microsoft Excel. To transform this raw data into insights, we use advanced analytics tools such as Python and R, which serve as a solid foundation for extracting insights. To store and manage this information, we rely on Azure Data Warehouse, which plays a key role in our data management. MLB teams need to allocate a budget for acquiring and implementing innovative technologies. To ensure successful execution and integration of these technologies into day-to-day operations, it is also essential to hire a Sports Scientist.

Before diving into this project, a thorough readiness assessment was performed to determine if the organization was prepared to take on this challenge. This assessment looked at the

organization's readiness for; data, team expertise and experience, analytical commitment, organizational and cultural change, and financial commitment. The organization exhibits strong readiness in data management, leveraging existing systems like FloQast, wearable technology, and Microsoft Excel. The project team possesses advanced analytics skills, particularly in Python and R, that are essential for extracting meaningful insights. While there is a commitment to analytics through predictive models and methodologies, fostering a culture that universally values data-driven decision-making will be a focal point. In terms of organizational and cultural change, there is moderate readiness, with acknowledgment of the need to optimize business processes. Yet, instilling a culture of continuous improvement may require targeted initiatives. Financial commitment is evident, with the allocated budget for innovative technologies and the addition of a Sports Scientist to the team. A comprehensive financial analysis, breaking down budgetary allocations and assessing the potential return on investment, will be crucial to ensure alignment with anticipated outcomes and long-term project sustainability.

The research project aims to enhance team performance and financial strategies. The expected outcomes include better injury prediction and prevention, leading to a reduced injury rate, improved athletic performance, and team success. With the help of data analytics, team officials can make informed decisions during player contract negotiations, aligning their investments with performance and risk. These outcomes result in MLB teams experiencing an increase in their revenue from ticket sales, merchandise, sponsorships, and fan engagement.

There are 4 sets of raw data being used to perform analysis and derive insights. This data has been retrieved from two sources, Fansgraphs.com and Rotowire.com, and verified against alternate data sources (ESPN). The data sets comprise of player injuries, player salaries, batter's stats, and pitcher's stats. As it was raw data, some data management and transformations needed to take place. For example, in the player injury data set there is a variable for the date a player was injured and the day that they returned to play. By subtracting Return Date from Injury Date, the Injury Duration

variable was created and used for further analysis. Another example is the transformation of the occurrence of a player injury into a numeric variable, where zero (0) represents no injury and one (1) represents a player suffering an injury. This data manipulation allowed for advanced statistical methodologies to be utilized with the raw data.

The analysis employs three key statistical methods: Linear Regression, revealing variables influencing injury likelihood; Applied Probability, assessing injury likelihood based on additional factors; and hypothesis testing, examining the impact of injury locations on player availability. Alternative approaches were considered but excluded. Logistic regression, suitable for binary variables, proved unfeasible due to non-percentage inputs. Sampling distribution, another option, was rejected as it involves separating players into different samples, offering limited insights in this context where individual player analysis is essential.

III. Results

The statistical methods discussed above helped derive valuable insights to be presented to MLB teams. To begin the aim was to identify which types of injuries were causing the most time away from play (Lower-Body vs. Upper-Body), as well as which position (Pitcher vs non-pitcher) experienced longer duration injuries. All injuries were classified as upper or lower based on the body positions listed in Appendix 2, and only injury types with more than 5 occurrences were included. Appendix 3 outlines total injuries by location after classification. To perform this analysis, hypothesis testing was performed using two t-tests. The null hypothesis of the first test states that the average duration of injury time for upper-body injuries is greater than the average duration for lower-body injuries.

The test returned average injury durations of 33.58 days for lower-body injuries and 49.98 days for upper-body injuries. These results suggest that we failed to reject the null hypothesis.

Because baseball is a sport where upper body extremities are more commonly utilized, this result makes sense. The second test was performed to test the null hypothesis stating that the average duration of pitcher injuries is greater than the average duration of non-pitcher injuries. This second test returned average injury durations of 29.1 days for non-pitcher injuries and 54.98 days for pitcher injuries. Like the first test, this result also suggests that we have failed to reject the null hypothesis. These results provide valuable insight, as we now have reason to believe that pitchers with upper body injuries are the most likely to be away from their team and therefore have a greater impact on a team. Figure 2 below displays the dispersion of player injury duration to give more context to our analysis.

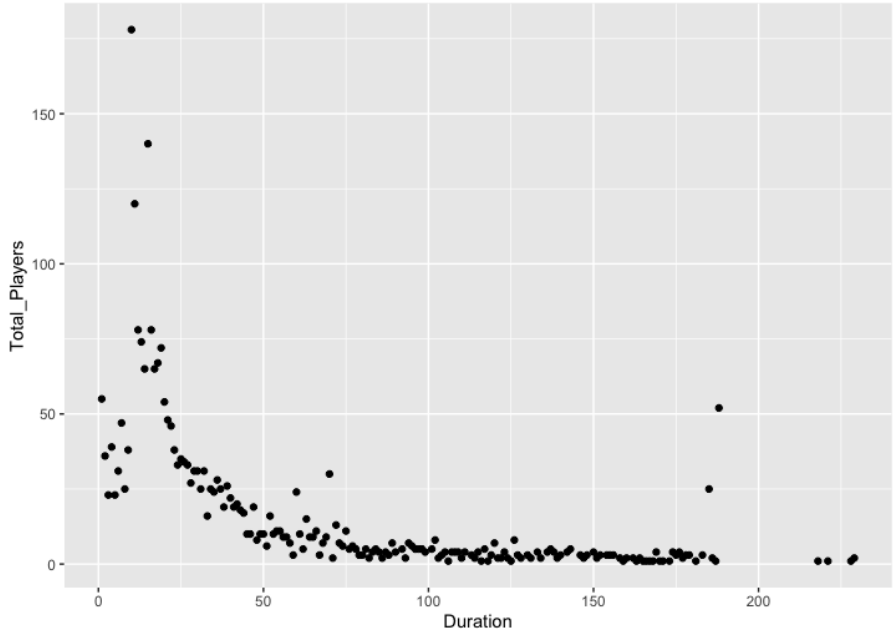


Figure 2: Scatter plot displaying the number of players (x) that experienced an injury at each duration length (y)

The study of applied probability has yielded valuable insights into the chances of injury and team accomplishment under specific circumstances. Table 1a shows that pitchers have an almost equal probability of being healthy or injured. While the results of our hypothesis testing (and the visual in Appendix 4) suggest that they may be more prone to injuries and may require longer recovery times, the results here indicates that pitchers have an equal chance of being healthy as they

do of being injured. Table 1b indicates that having a service time longer than the MLB average of 4.54 years does not significantly affect the occurrence rate of injuries.

Lastly, Table 1c investigates the correlation between team success and injuries, emphasizing the importance of injury prevention and management for team achievement. The data shows that if a team has more injuries than the average (22.4 injuries), the chances of missing the playoffs is 61%. This finding underscores the significance of injury management as a crucial factor in team success, along with other contributing factors. Appendix 5 displays the number of injured players across each MLB team over the 4 seasons examined in the dataset to support these findings.

Probability of Injury Given Position

	Non-Pitcher	Pitcher
Healthy	0.4943123	0.5056877
Injured	0.4665428	0.5334572

Probability of Injury Given Service Time

	Non-Veteran	Veteran
Healthy	0.6028956	0.3971044
Injured	0.5810409	0.4189591

Probability of Playoffs Given # of Injuries

	No	Yes
< Average	0.5606061	0.4393939
> Average	0.6111111	0.3888889

Table 1(a-c): Conditional Probability tables depicting the conditional probabilities of: a) player being injured given they are a pitcher b) a player being injured given they have been in the MLB longer than the average player (4.54 years) c) a team making the playoffs given they had more than the average number of injuries (22.4)

In investigating factors affecting injury likelihood, linear regression was used to explore variables directly influencing injury occurrence. A preliminary check of data variability was vital to confirm the method's appropriateness. Figure 3's box plot highlights extensive age dispersion among players in the 2020 season, justifying the inclusion of Age as a pertinent factor. Despite splitting the dataset into pitcher and batter individual datasets, and creating three linear models (as shown in Appendix 6), the models had limited effectiveness, indicated by low adjusted R-squared values. Introducing an interaction term (Age * Service Time) in the third model aimed to assess service

time's disproportionate impact on durability concerning age, but results were inconclusive due to a small coefficient. Examination of Batter and Pitcher datasets in Appendix 7 and 8 reinforced the finding that current variables are not robust injury indicators. Limited to the 2020-2023 seasons, the logical next step involves expanding the analysis to a broader dataset for refinement and validation. Integrating wearable technology readings from standardized player testing may provide more direct insights into injury-influencing factors, potentially revealing hidden patterns. This comprehensive approach ensures the analysis not only identifies limitations but also paves the way for further refinement and exploration toward a more robust injury prediction model.

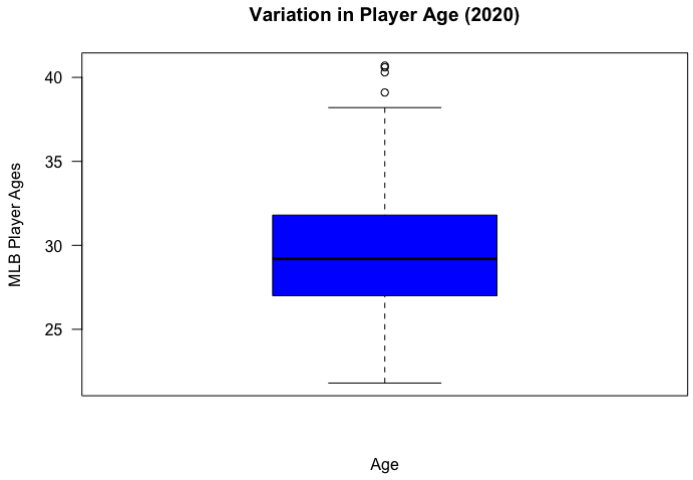


Figure 3: Box plot highlighting variation in age across MLB player in the 2020 season

IV. Conclusion

This project uses predictive analytics to help MLB organizations manage injuries, mitigate financial risks, and optimize team productivity. The initiative integrates injury data, player demographics, and financial metrics to provide decision-makers with valuable tools. The objective is to identify injury predictors and contribute to a culture of data-driven decision-making. Enhancements to the project include incorporating more vast data and exploring wearable technology readings to improve the injury prediction model.

Statistical methods such as Linear Regression, Applied Probability, and Hypothesis Testing provided valuable insights. The analysis of injury durations highlights the significant impact of upper-body injuries on Pitchers. Applied probability helps to understand the correlation between team success and injuries, emphasizing the fact that teams with above-average injury occurrences are at a greater risk of missing the playoffs. Therefore, there is a need for proactive injury prevention strategies to ensure teams can perform at their best.

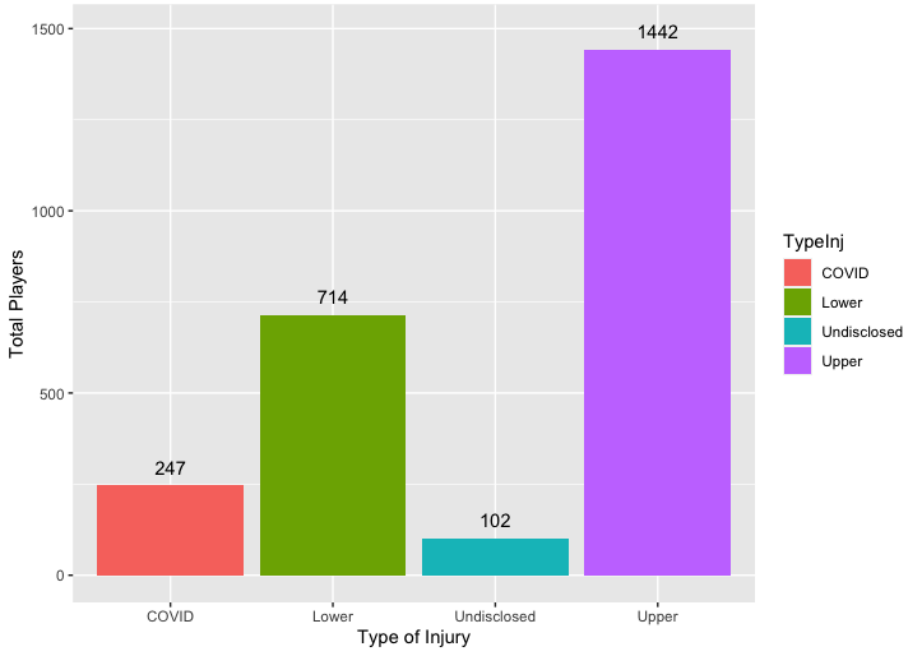
In conclusion, this BI project transcends statistical analyses, serving as a strategic tool for MLB organizations to navigate the unpredictable landscape of player injuries. The insights derived not only refine decision-making processes but also contribute to fostering a culture of continuous improvement, ensuring sustained team productivity and financial resilience in the fiercely competitive MLB arena.

Appendices

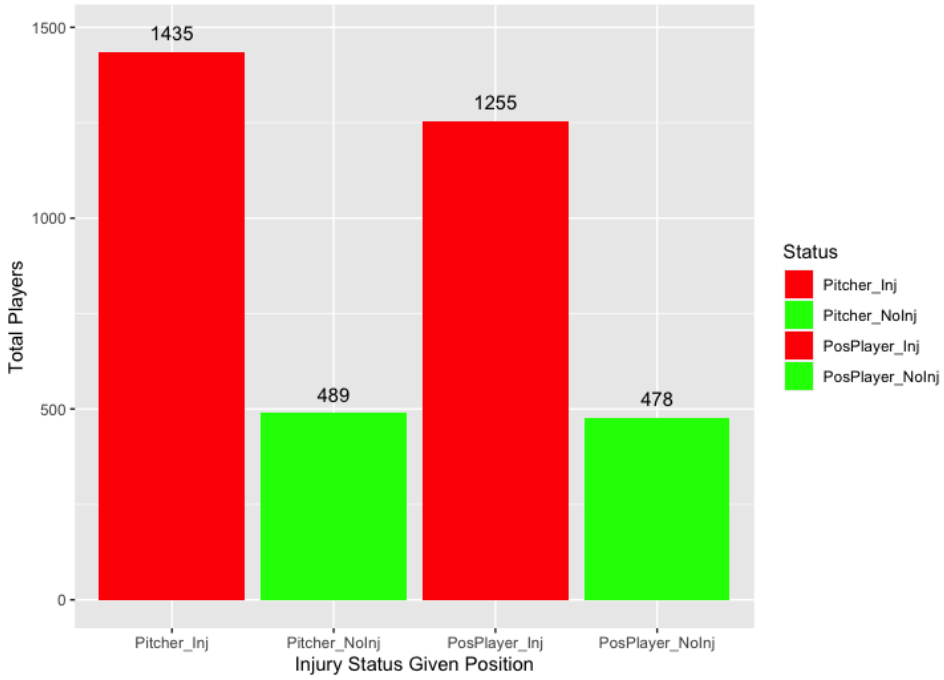
Appendix 1: Business Intelligence Roadmap

Upper	Lower
Abdominal	Toe
Back	Ankle
Bicep	Foot
Elbow	Achilles
Finger	Calf
Neck	Knee
Wrist	Hip
Tricep	Plantar
Tommy John	Groin
Thumb	Hamstring
Thoracic	Quad
Shoulder	Heel
Oblique	
Lat	
Intercostal	
Hand	
Pec	
Concussion	

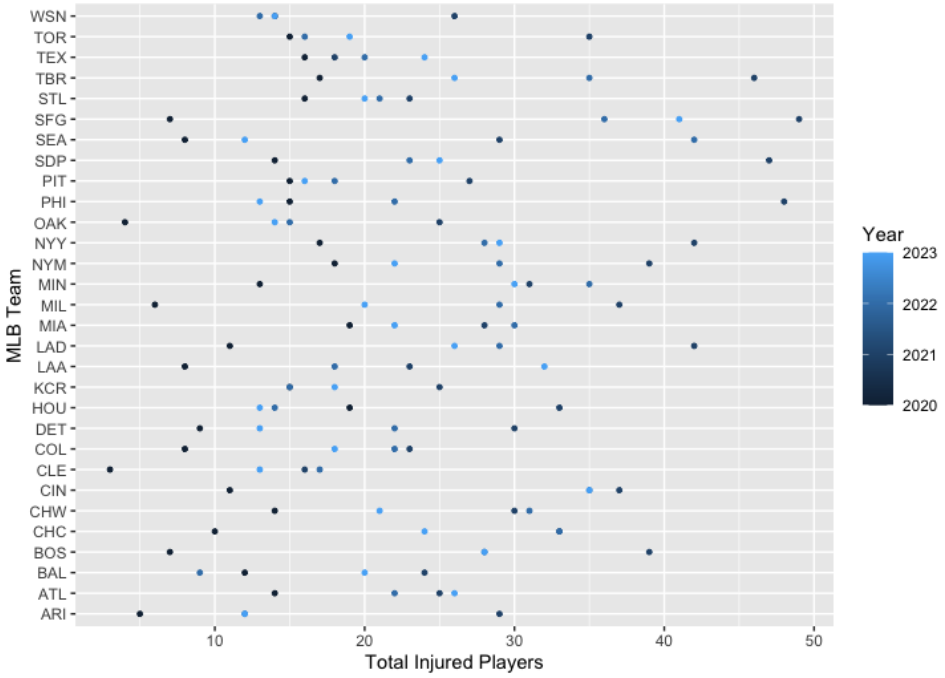
Appendix 2: Upper & Lower Body injury locations



Appendix 3: Total players injured by injury type



Appendix 4: Counts of player status (healthy vs injured) based on position



Appendix 5: Number of injured players across each MLB team over 2020-2023 seasons

Linear Regression Models (All Player Injuries)

Variable	Model 1	Model 2	Model 3
Intercept	0.506	0.536	0.554
Age	0.008	0.007	0.006
Service Time	-0.002	0.001	-0.004
Pitcher	0.016	0.014	0.014
Salary	NA	$2.2 * 10^{-9}$	$2.1 * 10^{-9}$
Age * Service Time	NA	NA	0.0002
p value	0.017	0.011	0.022
Adjusted R-squared	0.002	0.0025	0.0022

Appendix 6: Results of linear regression depicting the coefficients of each variable included in the model as well as associated p values and Adjusted R-squared (All data)

Linear Regression Models (Pitcher Injuries)

Variable	Model 1	Model 2	Model 3
Intercept	0.547	0.532	0.568
Age	0.007	0.007	0.007
Service Time	0.005	0.004	0.005
Games	0.0003	0.006	0.006
Games Started	0.002	0.028	0.027
Innings Pitched	NA	-0.006	-0.005
Playoffs	NA	NA	-0.06
p value	0.0003	$1.9 * 10^{-8}$	$1.27 * 10^{-9}$
Adjusted R-squared	0.011	0.027	0.032

Appendix 7: Results of linear regression depicting the coefficients of each variable included in the model as well as associated p values and Adjusted R-squared (Pitchers only)

Linear Regression Models (Batters Injuries)

Variable	Model 1	Model 2	Model 3
Intercept	0.606	0.602	0.658
Age	0.001	0.001	$3.01 * 10^{-5}$
Service Time	0.012	0.013	0.014
Games	0.005	0.005	0.004
Plate Appearances	-0.001	-0.001	-0.001
Stolen Bases	NA	0.002	0.0002
Playoffs	NA	NA	-0.07
p value	$4.81 * 10^{-9}$	$1.21 * 10^{-8}$	$5.3*10^{-10}$
Adjusted R-squared	0.031	0.031	0.037

Appendix 8: Results of linear regression depicting the coefficients of each variable included in the model as well as associated p values and Adjusted R-squared (Batters only)