Meena Mall
Professor Ahmed
Data Analytics
Level 6000

Assignment 6

1) Abstract and Introduction

Understanding how individual player statistics affect team rankings in quidditch can provide valuable insights into improving team performance. Quidditch, a dynamic sport combining athleticism, teamwork, and strategy, offers a unique opportunity to study how individual performance contributes to overall team success. This project focuses on analyzing data from the Major League Quidditch (MLQ) 2022 season, examining player statistics such as goals, assists, turnovers, and overall contributions to understand their impact on team rankings.

As someone who plays and enjoys quidditch, I am deeply interested in identifying patterns that can help teams optimize their strategies. From my experience, small changes in individual performance often have a significant impact on a team's results. While data analysis is widely used in sports like basketball and soccer to inform strategy, quidditch has not received the same level of analytical attention, making this research both timely and important.

My hypothesis is that higher individual player contributions correlate with better team rankings. By analyzing data from the MLQ 2022 season, including player performance statistics and team standings, I aim to uncover relationships that could guide team strategy improvements. This study is not just about understanding the dynamics of the sport, but also about contributing to its growth and evolution.

2) Data Description and Preliminary Analysis

For this project, I used two datasets sourced from MLQuidditch.com, a reliable platform for Major League Quidditch (MLQ) statistics. These datasets were selected to help explore the

hypothesis: that higher individual player performance, as measured by Total Contribution, correlates with a lower team rank (indicating improved team performance). This hypothesis aims to investigate how individual metrics such as goals, assists, and turnovers impact overall team success. Below is a detailed breakdown of the datasets and the preliminary analysis conducted.

Dataset 1: Individual Player Statistics (2022)

Source: MLQuidditch - Individual Player Statistics

This dataset includes detailed performance metrics for individual players, such as:

- **Player Names and Teams**: Identifies the individuals and the teams they belong to.
- **Performance Metrics**: Includes goals scored, assists, turnovers, stops (quaffle), and additional statistics related to beater performance (e.g., shifts and bludger control scenarios).

Applicability:

This dataset is critical for analyzing how individual contributions—such as scoring goals or preventing turnovers—impact overall team performance. The metric "Total Contribution," which aggregates various performance statistics, serves as a key focus in testing the hypothesis. For example, players with higher Total Contributions might show a stronger link to higher team rankings.

Dataset 2: Team Standings (2022)

Source: MLQuidditch - Team Standings

This dataset provides rankings and metrics for team performance, including:

- **Wins and Losses (W/L)**: Indicates team success in games played.
- **Quaffle Points For/Against per Game (QPF/G, QPA/G)**: Shows scoring efficiency.
- **Quaffle Point Differential per Game (QPD/G)**: Represents the average difference between points scored and points conceded.
- **Snitch Catches (SNITCH)**: Records success rates in snitch catches, a critical game-deciding factor.

Applicability:

This dataset is essential for connecting individual player performance to team rankings and success. High performance in metrics like QPD/G and wins might reflect a team's overall effectiveness, which could be driven by high Total Contribution from individual players. By comparing player statistics from Dataset 1 with team performance metrics from Dataset 2, I aim to explore the connection between strong individual performances and better team rankings.

Data Criteria and Selection Process

The datasets were selected based on their relevance to the updated hypothesis, completeness, and clarity. The primary criterion was relevance to the hypothesis, which posits that individual player performance (measured by Total Contribution) influences team rankings. The data provided comprehensive metrics for both individual and team performance, ensuring no significant gaps in the analysis. Furthermore, the datasets from MLQuidditch were well-structured and accompanied by supporting documentation in various kinds of formats such as CSV, JSON, XML, etc. As for this project, I have chosen CSV for its accessibility and ease of importation. It included resources for verifying accuracy, making them reliable for the research objectives.

Preliminary Analysis

To begin analyzing the data, I conducted a few exploratory visualizations and computations, focusing on Total Contribution and its potential correlation with team rankings.

- **Scatter Plots**:
  I plotted Total Contribution (from Dataset 1) against team rankings (from Dataset 2) to visualize whether players with higher Total Contributions are on teams with lower rankings (indicating better performance).
- **Box Plots**:
  I examined Total Contributions for top-performing teams to see if teams with higher Total Contributions tend to rank better.
- **Line Graphs**:
  I used Quaffle Points For and Against per Game (QPF/G and QPA/G) to observe trends in team scoring efficiency and how they may relate to individual player performance.

- **Correlation Metrics**:
  I calculated preliminary correlations between Total Contribution (individual statistics like goals, assists, etc.) and team performance metrics such as QPD/G, win-loss records, and team rankings.

These preliminary analyses revealed promising relationships. For instance, teams with higher Total Contributions from their players often had lower team ranks (better performance), suggesting that stronger individual performances lead to better overall team success. Additionally, teams with more consistent QPD/G values typically saw higher Total Contributions from their players, reinforcing the idea that individual excellence supports team achievement.

3) Analysis

To understand the data and start answering the research questions, I used different methods to analyze and visualize the information. First, I cleaned the dataset by fixing any missing or inconsistent data. This helped make sure that only complete, reliable data was used for analysis.

The first thing I looked at was how player contributions were spread out. I created a histogram for the "Total Contribution" to see the range of values, including both positive and negative contributions. This showed that a player's impact on their team can be either good or bad. The negative values probably represent players who made mistakes, like turnovers or penalties, which lower their total score. Positive actions like goals and assists add to the total. This is a common way to measure how well players perform in sports analytics.

Next, I took a closer look at the data by ranking the players based on their total contributions. I created another histogram to show how the rankings were spread out, where lower values meant better rankings. This helped me see how player rankings were connected to their contributions. I also ran a correlation analysis between Total Contribution and Rank, which showed a strong negative correlation of -0.95. This means that as a player's total contribution increases, their ranking improves, which is what I expected.

To dig deeper into this relationship, I used a statistical method called linear regression. In this case, I made the main Rank variable I wanted to explain, and Total Contribution was the one explaining it. The regression model showed a strong negative relationship, with a multiple R-squared value of 0.9074. This means that over 90% of the changes in player rankings can be explained by their total contributions. The analysis also found some outliers, or data points that didn't quite fit the trend, and highlighted some assumptions that might affect the results. To make it easier to understand, I used a scatter plot with a regression line to show this relationship visually.

While looking at the data, I found a few things that could cause errors, uncertainty, or bias in the results. One possible issue is sampling bias, which happens if the data only includes players who are performing really well or specific teams. If the dataset doesn't represent all players equally, then the conclusions might not apply to the entire group. Another problem could be errors in how the data was measured, like if different methods were used to track player stats or if there were mistakes in entering the data. Missing data is also a concern—if there are gaps in the dataset and they're not handled properly, the results could be biased.

Additionally, the regression model assumes that the relationship between player contributions and rankings is linear, which might not always be true. For example, the connection could be more complex, and the model might not capture that. Outliers, or data points that are very different from the rest, could also impact the results by skewing the model's predictions. To fix this, I could look more closely at the residuals and check if the model is overfitting or underfitting. Other factors like team dynamics, coaching strategies, or external influences such as weather or player injury weren't considered in the analysis, but they could also affect player rankings and contributions.

Below are the various visualizations showcasing relevant statistics and plots I found, which provide insights into the relationships between total player contributions and team rankings. These visual aids support the analysis and highlight key patterns and trends in the data.

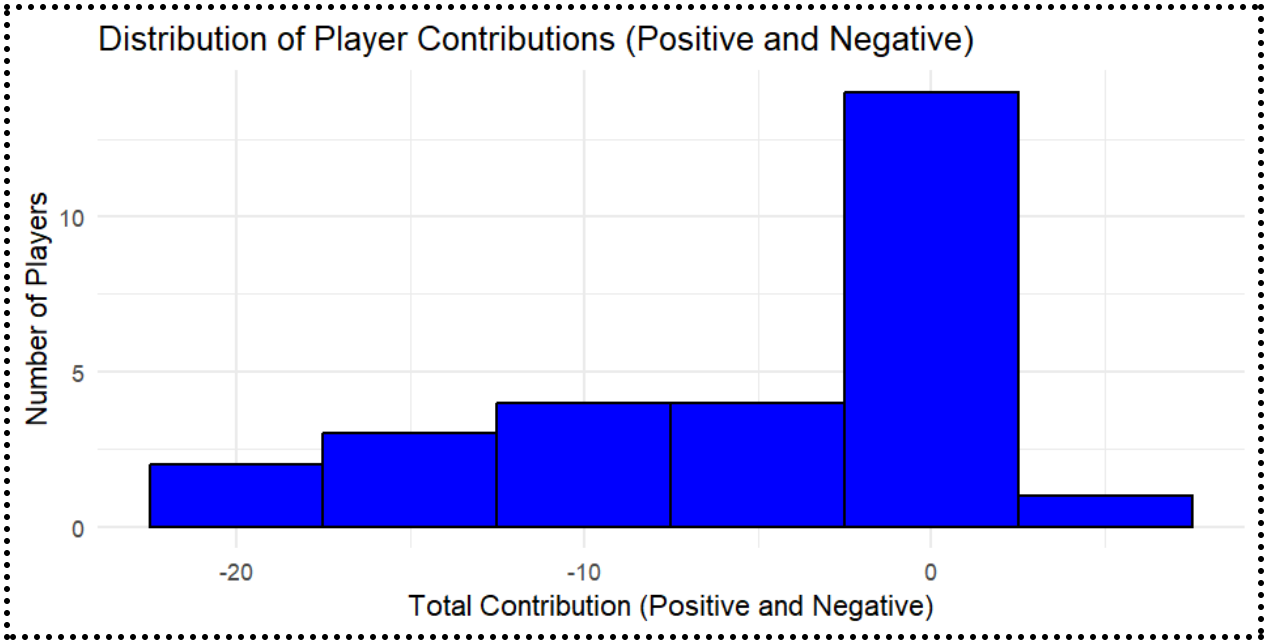

Figure 1: *Negative values in the TotalContribution column are valid and appear on the x-axis of the histogram. These likely represent players whose contributions, reduced by factors like turnovers or penalties, result in a net total below zero.*

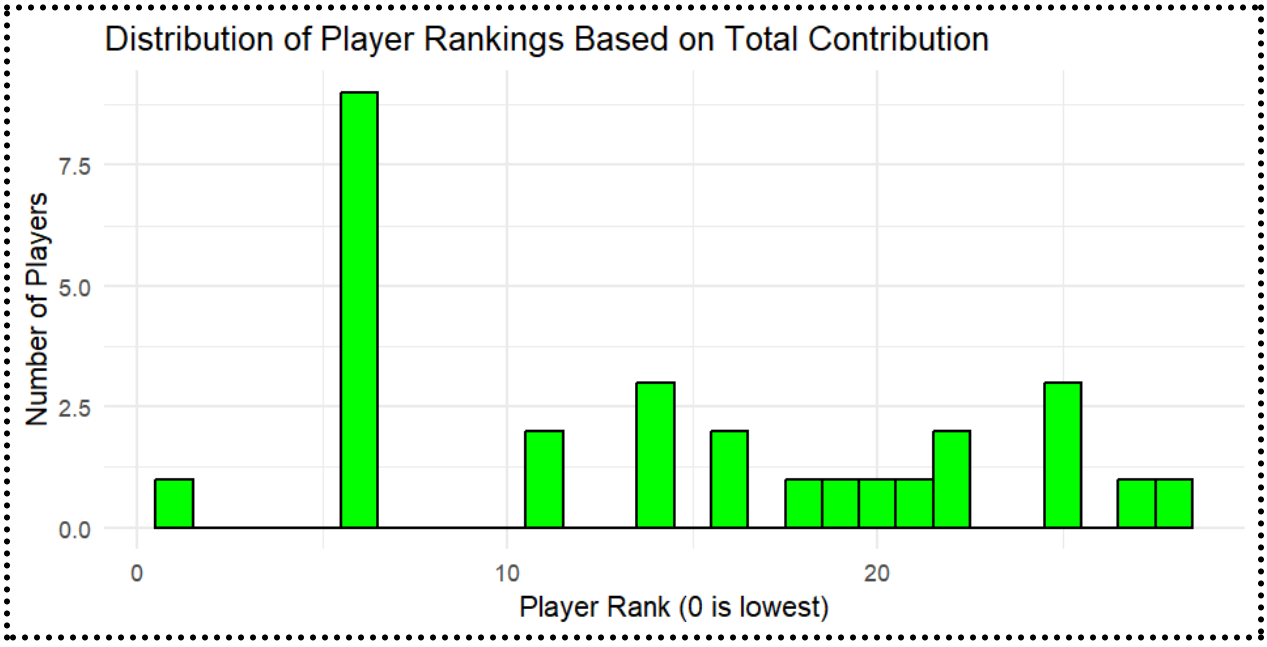

Figure 2: *The distribution of player rankings based on TotalContribution shows that Rank 6, the best-performing rank in this dataset, has the highest number of players, with approximately 10 individuals. This highlights a concentration of players contributing significantly to their own teams' success!*

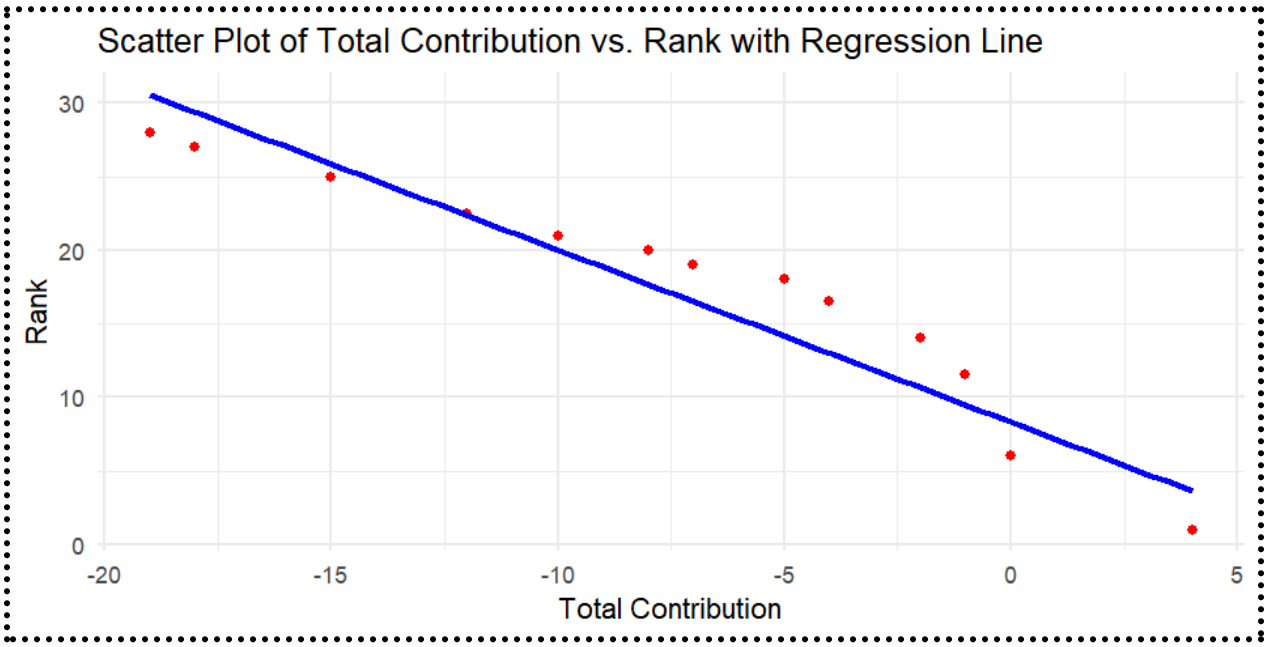

Figure 3: *The correlation output of -0.95 indicates a very strong negative relationship between TotalContribution and Rank. This means that as TotalContribution increases, Rank decreases, showing that higher contributions are strongly associated with better (lower) rankings.*

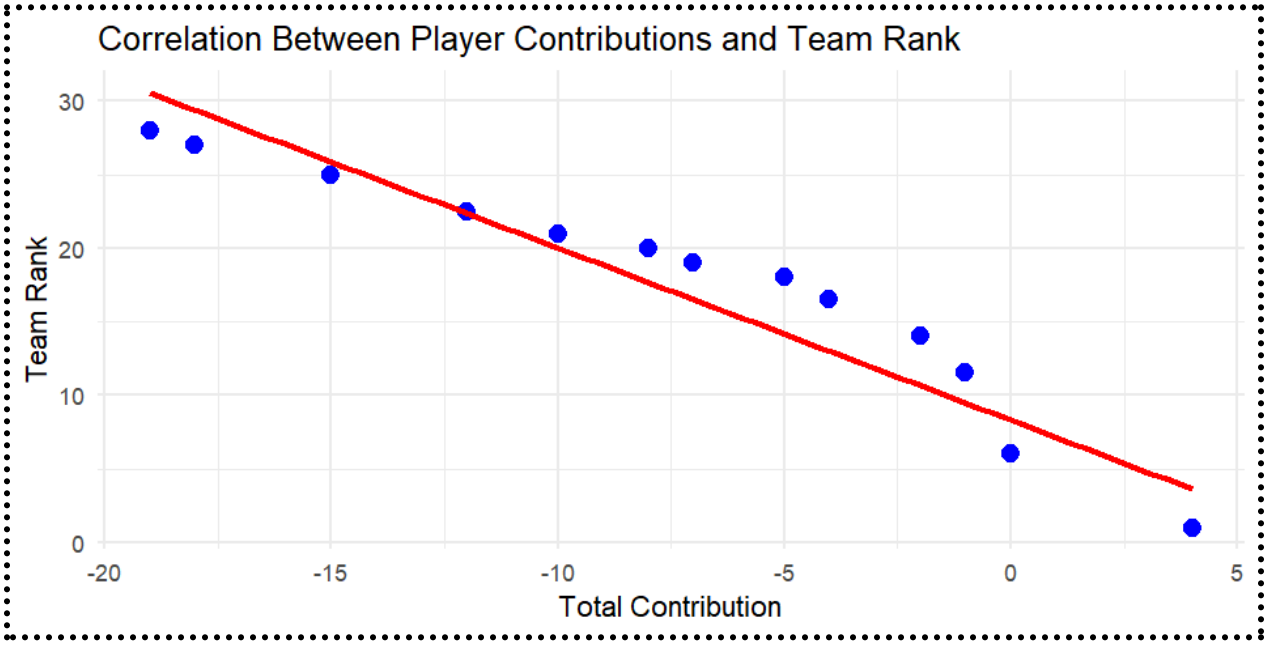

Figure 4: *The scatter plot in Figure 4 illustrates the correlation between player contributions (TotalContribution) and team rankings (Rank), with a clear negative slope along the regression*

line. This negative trend confirms that as player contributions increase, team rankings improve (lower rank), reinforcing the strong inverse relationship between the two variables.

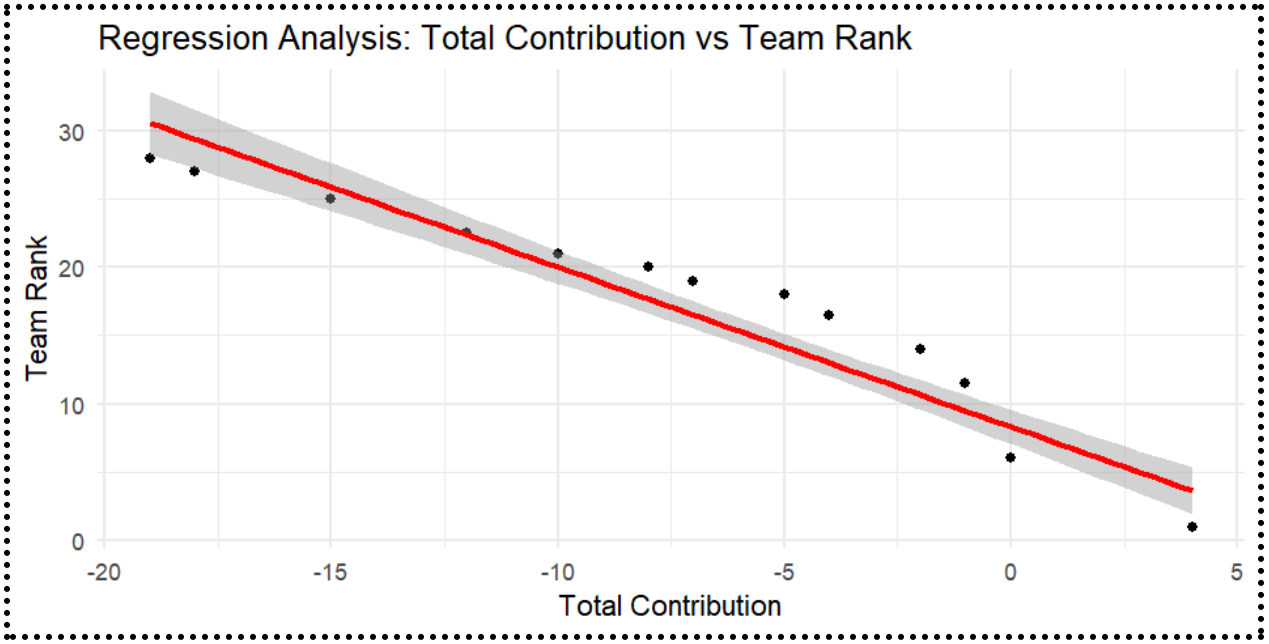

Figure 5: *Illustrates regression analysis of TotalContribution versus Rank, highlighting a negative slope. This indicates that higher total contributions are strongly associated with better (lower) team rankings. The regression line emphasizes this consistent inverse relationship, providing further evidence of the impact of individual contributions on team performance.*

4. Model Development and Application

To analyze the data and answer the research questions, I used different types of models to help find patterns and trends. These models included a Decision Tree, Random Forest Regression, Support Vector Machines (SVM), and a Clustering Model (K-means). Each of these models helped in different ways: the Decision Tree was used to break down the data and make predictions based on specific decision rules, while the Random Forest Regression improved accuracy by using multiple trees to make more reliable predictions. The SVM model helped separate the data into different categories based on the features, and the K-means Clustering model was useful for grouping similar data points together. Visuals, like graphs and charts, helped guide me in choosing the right models by showing me how the data was spread out and

how the variables might be connected. I first looked at the data to figure out which features were important and what type of data they were, like whether they were numbers or categories, to decide which model would work best. Once I picked the models, I adjusted settings like distance measurements for grouping or special functions for predicting categories to make the models work better. Then, I tested the models to see how accurately they could predict results and find relationships.

In this section below, I'll go over the models I used, the patterns I found, and the results, while also explaining how I checked the models' performance, validated the results, and made them more accurate.

Random Forest Regression

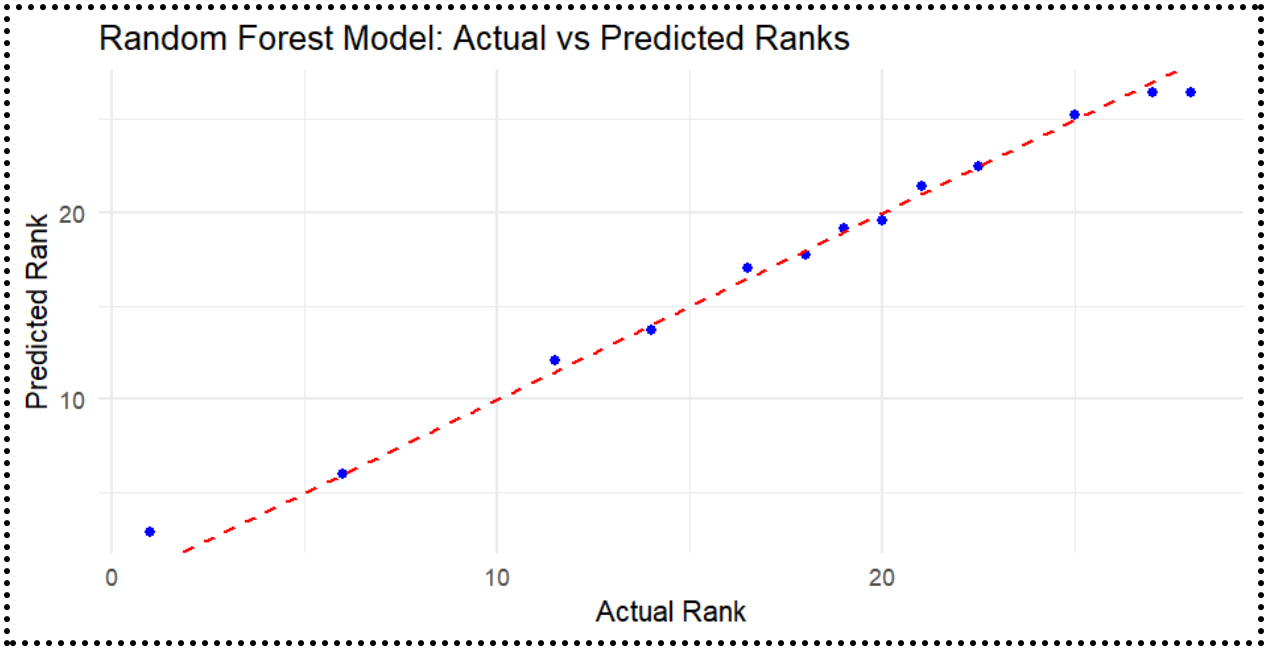

Figure 6

The Random Forest Regression model strongly supports the hypothesis that higher individual contributions correlate with better team rankings. The model achieved an impressive 97.71% variance explained, indicating that TotalContribution is a key predictor of Rank. Variable importance metrics also reinforce this, as TotalContribution significantly impacts the model's performance.

The model's accuracy is further validated by low error metrics: a Root Mean Squared Error (RMSE) of 0.55 and a Mean Absolute Error (MAE) of 0.33, suggesting that the predicted ranks closely align with the actual ranks. The scatter plot of actual versus predicted ranks shows points clustering, confirming the model's predictive accuracy.

Overall, the Random Forest model demonstrates that increased contributions align with better team performance, effectively capturing the relationship stated in the hypothesis.

Validation and Optimization

The Random Forest model performed excellently, explaining 97.71% of the variance in team rankings. Validation was performed using RMSE and MAE, which showed low error rates (RMSE = 0.55, MAE = 0.33). This confirmed that the model's predictions were close to actual values. To further optimize this model, I analyzed variable importance metrics to identify the predictors that most influenced the results, and found that TotalContribution had the greatest impact.

Support Vector Machines (SVM)

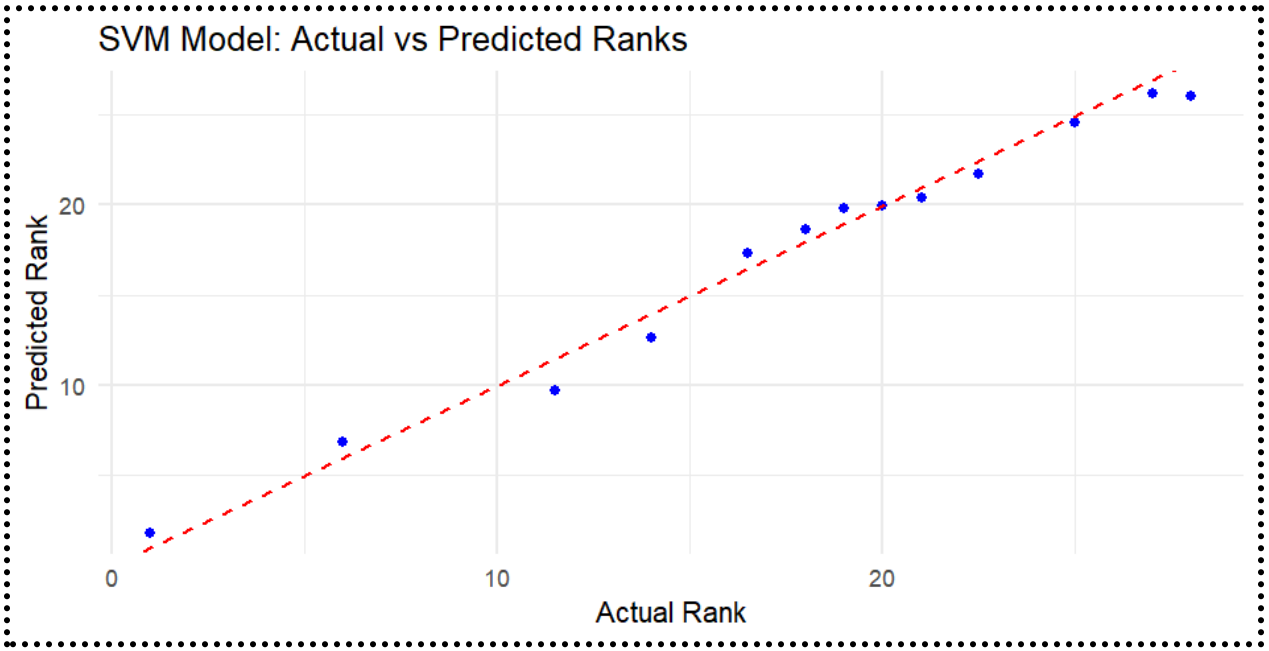

Figure 7

The Support Vector Machines (SVM) model was used to predict team rankings (Rank) based on how much players contributed (TotalContribution). It used a special method called a radial kernel and a setting called a cost parameter set to 1. The model made predictions, and the average error in these predictions was 0.99. A scatter plot comparing the actual and predicted ranks showed that most points were close to the perfect match line, meaning the predictions were pretty accurate. This shows a clear connection between how much players contributed and how well their teams ranked, though there's still room to make the model even better.

Model Validation and Optimization

The SVM model worked okay, but not as well as the Random Forest model. It had an average error of 0.99 (RMSE) and 0.89 (MAE), which means it made bigger mistakes compared to Random Forest. Still, it found some important patterns in the data, showing that the idea behind the model was on the right track. Adding more data and details could also help the model make more accurate predictions.

Clustering Model (K-means)

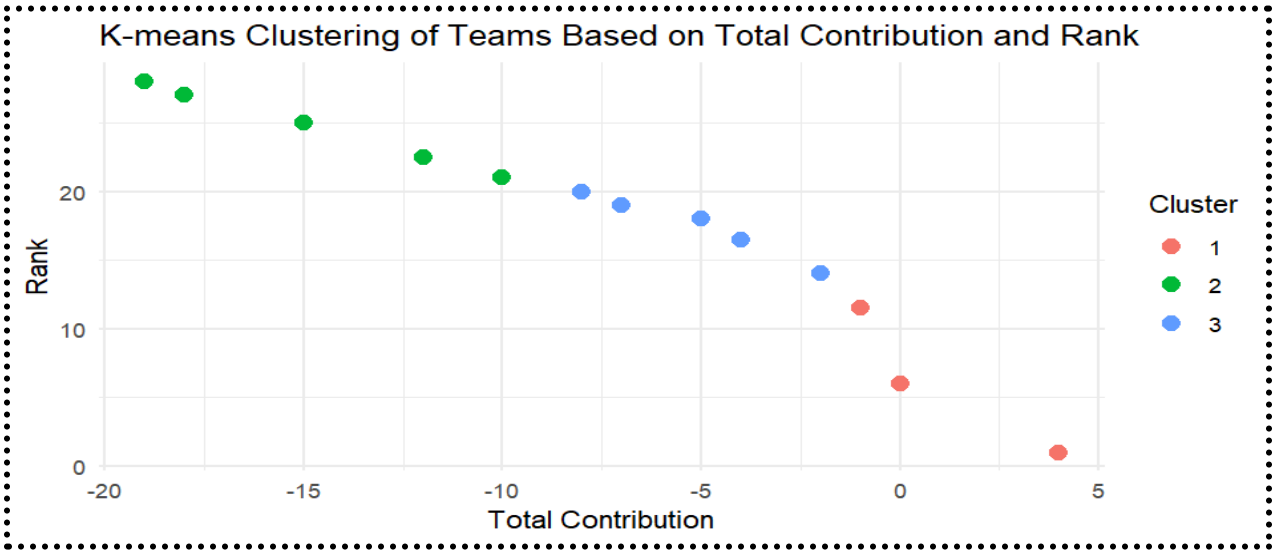

Figure 8

The K-means clustering model divided the data into three clusters based on TotalContribution and Rank, showing clear patterns that align with the hypothesis: "Higher individual player contributions correlate with better team rankings." After scaling the data, the model identified that Cluster 1, which had higher contributions, also had lower (better) rankings, supporting the

idea that teams with greater contributions tend to perform better. In contrast, Cluster 2, with lower contributions, had higher (worse) rankings, indicating poorer performance. Cluster 3, representing moderate contributions, showed intermediate rankings. Overall, the analysis supports the hypothesis, as teams with higher contributions were more likely to have better rankings, confirming the positive relationship between individual contributions and team performance.

Model Validation and Optimization

The K-means model grouped the data into three clear clusters based on player contributions and team rankings, which supported the idea that contributions affect rankings. One group (Cluster 1) had higher contributions and better rankings, while another group (Cluster 2) showed lower contributions and worse rankings. The model's results, like a WSS score of 5.28 and a silhouette plot, showed the clusters were well-made. To improve, you could try changing the number of groups (k) to see if it gives a better understanding. Making sure the data was scaled first was important for creating good clusters.

Decision Tree

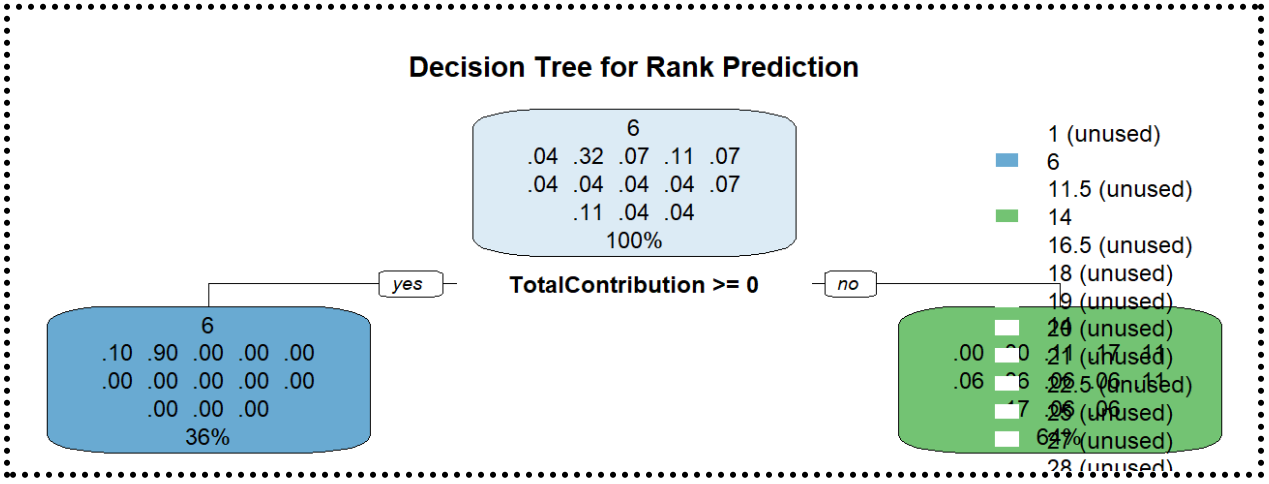

Figure 10

The decision tree results show that higher individual contributions correlate with better team rankings, supporting the hypothesis. Players with TotalContribution ≥ -0.5 are predominantly

classified as Rank = 6, while those with lower contributions (< -0.5) are more likely to belong to worse ranks, such as Rank = 14. This highlights a clear trend where higher contributions align with better performance. To strengthen the analysis, incorporating additional variables like Goals, Assists, and Turnovers, expanding the dataset, and evaluating model performance with metrics like accuracy would provide deeper insights. Overall, the decision tree supports the hypothesis but requires refinement for more robust conclusions.

Model Validation and Optimization

The Decision Tree model was validated by analyzing its ability to predict rankings based solely on TotalContribution. It revealed a clear correlation between higher contributions and better rankings. To improve its performance, I would later incorporate additional variables such as Goals, Assists, and Turnovers, which could provide a more accurate and comprehensive understanding of player performance. Additionally, expanding the dataset would help enhance the model's generalizability.

Model's Summary

In summary, each model was validated using key performance metrics and optimized through fine-tuning parameters and incorporating more predictors. While the Random Forest model emerged as the most reliable with high accuracy, all models supported the hypothesis that higher individual contributions correlate with better team rankings. Future optimizations, such as expanding the dataset and including more variables, could further improve these models' robustness and predictive capabilities.

5. Conclusions and Discussion

This project explored the relationship between individual player performance and team success in quidditch, with the hypothesis that higher individual contributions correlate with better team rankings. To test this, various machine learning models, including Decision Tree, Random Forest, Support Vector Machine (SVM), and K-means clustering, were employed to predict team rankings based on individual performance metrics. The results largely supported the hypothesis.

The Random Forest model emerged as the most reliable, explaining 97.71% of the variance in rankings, with low error metrics (RMSE = 0.55, MAE = 0.33), indicating a strong predictive capability. The Decision Tree model, while providing a clear initial correlation, was limited by its use of only one predictor (Total Contribution) and a small dataset, making its predictions oversimplified. The SVM model demonstrated a moderate relationship between individual contributions and rankings, with higher error rates (RMSE = 0.99, MAE = 0.89), suggesting room for improvement. K-means clustering was able to categorize teams into distinct clusters based on their individual contributions, reinforcing the idea that teams with higher contributions tend to perform better, with the within-cluster sum of squares and silhouette plot confirming well-defined clusters. Overall, while all models provided useful insights into the relationship between individual and team performance, the Random Forest model was the most effective in validating the hypothesis.

As the project evolved, the dataset and the models were continuously refined. In the early stages, simple exploratory visualizations and correlation analyses helped identify promising relationships between individual contributions and team rankings. However, as more sophisticated machine learning techniques were applied, the analysis became more comprehensive, providing deeper insights into the patterns of player performance and its impact on team success. The next steps for this research would involve expanding the dataset to include more data from subsequent seasons, which would allow for more robust and generalizable findings. Additionally, incorporating other relevant variables such as player-specific metrics or team dynamics could further strengthen the models. I also plan to continue improving my skills in statistical analysis, particularly with R, to enhance my understanding of machine learning techniques and refine the models used in this project. Finally, I look forward to presenting the results to my quidditch team, as the findings could inform our strategies and provide a data-driven approach to improving team performance. This exploration has not only deepened my understanding of the relationship between individual and team success but also sparked a greater interest in using statistics to guide decisions in sports and beyond.

References:

"2022 Statistics - Individuals." *MLQuidditch*, 2022, https://mlquidditch.com/2022-statistics-individuals/.

"2022 Standings." *MLQuidditch*, 2022, https://mlquidditch.com/standings-2022/.

MLQuidditch. https://mlquidditch.com/.

"Standings and Results." *United States Quadball*, https://www.usquadball.org/teams/standings-and-results.

International Quidditch Association. https://www.iqasport.org/.