MGSC661 Final Project Report

December 11, 2023

1 Introduction

In the dynamic field of sports analytics, the Olympic Games offer a rich and multi-faceted dataset for examining the various elements that influence athletic performance. My study utilizes an expansive dataset that captures a wide range of variables, including "ID," "Name," "Sex," "Age," "Height," "Weight," "Team," "NOC" (National Olympic Committees), "Games," "Year," "Season," "City," "Sport," "Event," and "Medal."

A key focus of this project is to assess the impact of the host city on the performance of athletes from the host nation. This aspect is particularly intriguing, hypothesizing that hosting the Olympics might provide a competitive edge to local athletes. To explore this hypothesis, I employ sophisticated feature engineering techniques. These techniques allow me to transform raw data into more meaningful variables, providing a deeper and more accurate analysis of the host city's influence on athletes' performance.

Additionally, this project aims to develop an advanced predictive model to forecast future Olympic medal outcomes for participating countries. This model, enriched by the insights gained through feature engineering, not only takes into account conventional metrics of athletic performance and demographics but also incorporates the host city factor as a critical variable. This integration offers a more comprehensive understanding of the factors that drive Olympic success.

Through this dual approach, combining feature engineering with predictive modeling, my study seeks to make a significant contribution to the field of sports analytics. The findings are anticipated to be of immense value to athletes, coaches, sports strategists, and enthusiasts, providing new perspectives on performance optimization and strategic planning in the context of the Olympic Games.

2 Data Description

2.1 Data Cleaning

Data preparation is a critical step in ensuring the accuracy and relevance of my analysis. The initial dataset, consisting of 271,116 records, presented a significant challenge with numerous missing values in key variables such as height and weight. To address this, I decided to remove all entries with missing data in these fields, ensuring the integrity and consistency of the dataset.

For the purpose of analyzing the impact of the host city on athlete performance, my approach shifted from an individual to a country-level analysis. This change allows me to evaluate the overall performance impact on athletes from host countries, rather than focusing on individual athlete outcomes. This perspective is crucial to effectively discern patterns and trends that are indicative of the host city's influence.

I introduced a new binary variable, "Comes_from_City," to indicate whether an athlete's country of origin matched the host city's country (labelled as 1) or not (labelled as 0). This was achieved by matching the "City" variable with the corresponding "NOC" entries. The classification of the "Sport" variable into three categories – "Individual," "Team," or "Both" – was another strategic decision to refine the analysis based on the type of sport. This classification aids in understanding how different types of sports might be influenced by the host city factor.

In my Python-based data processing, I grouped the athlete data by "Year" and "NOC", performing several aggregation operations. This is because time is also important for predicting future performance. These included counting the instances of "Comes_from_City", segregating counts by "Sex," and calculating the averages for "Height," "Weight," and "Age." I also renamed the "ID" variable to "Number_of_Athlete" to reflect the count of athletes, and aggregated counts for each "Sport_Type," as well as counts of female and male athletes.

A crucial aspect of the data preparation was the quantification of the medals. I adopted the 4:2:1 points system proposed by Jeff Z. Klein in a New York Times blog post. This system, a compromise between the total-medals and golds-first methods, assigns points to each medal type (4 points for gold, 2 for silver, 1 for bronze), allowing me to create a composite "Total_Medal_Score" for each country and year. This metric provides a nuanced view of a country's performance in the Olympics, considering not just the number of medals won but their type as well. This comprehensive approach to data preparation sets the stage for a more accurate and insightful analysis of the impact of the host city on Olympic performances.

In further refining my dataset for a country-level analysis, I dropped the variables "Name" and "Team." The rationale behind omitting the "Name" variable was that individual athlete identities are not essential for this analysis, as the focus is on aggregate country performance rather than individual achievements. Similarly, the "Team" variable was removed because it largely duplicated the information provided by the "NOC" variable. After completing the data preparation and cleaning process, the final dataset was significantly refined, comprising 3,085 records across 13 variables.

2.2 Data Exploration

I conducted an exploration of both the original and the cleaned Olympic datasets. The original dataset, comprising 271,116 records, was first subjected to a process of removing all entries with missing values. This step reduced the dataset to 206,165 records. A modification I made was in the "Medal" column, where I replaced all NA values with "No Medal." This allowed for a more accurate representation in subsequent visualizations.

I then created a histogram to visualize the distribution of medals. This histogram revealed a significantly higher count of "No Medal" entries compared to the actual medals – "Bronze," "Gold," and "Silver." Excluding "No Medal," both "Bronze" and "Gold"

medals hovered around the 10,000 mark, while "Silver" was slightly less common. An interesting pattern emerged when I plotted this histogram as a stacked bar chart, segmented by the "Comes_from_City" variable. It showed that athletes from the host city's country had a higher count of Gold medals compared to Bronze and Silver, suggesting a potential correlation between host city advantage and athlete performance.

Further, I analyzed the Medal versus Year data as a stacked bar chart, color-coded by medal type. This visualization highlighted an increasing trend in the total number of medals over the years. Notably, post-1994, the total medal count fluctuated, alternating between significant increases and decreases every other year. Additionally, the rate of increase in Bronze medals was more pronounced than that for Gold and Silver.

For my cleaned dataset, I focused on pair plots to examine correlations. These plots indicated a generally positive relationship between "Comes_from_City" and "Total_Medal_Score," as well as between "Comes_from_City" and both "Sport_Type" and "Number_of_Athlete." This suggests a trend where a greater number of athletes from the host city's country participated and were successful in the Games, supporting the hypothesis of a host city advantage. This phase of data exploration was instrumental in shaping the direction of my further analysis, providing initial insights into the complex dynamics of Olympic performance and the potential influence of the host city.

3 Model Selection

3.1 Data Preprocessing

In the preprocessing phase, I tailored my approach to align with the characteristics of the selected model, Random Forest. Given that Random Forest models are generally robust to outliers, I decided not to perform outlier testing, which streamlined the preprocessing steps.

The first significant preprocessing task I undertook was MinMax Scaling. Considering the presence of categorical variables in my dataset, MinMax Scaling was chosen as it effectively normalizes the data, maintaining the structure of the dataset while scaling numerical values to a standard range. This process was crucial, particularly for variables that spanned a wide range of values. The dataset post-scaling was labeled as the 'scaled dataset.'

Subsequently, I applied Principal Component Analysis (PCA) for dimensionality reduction. The scree plot was an essential tool in this process, helping me identify the number of components that explained a significant portion of the variance in the data. The plot indicated that six components corresponded to the 'elbow,' a point beyond which the marginal gain in explained variance diminishes significantly. Additionally, the first six components cumulatively accounted for over 80% of the variance, conforming to the 80% rule used in PCA. Despite these findings, I opted to retain all 12 components for this stage of the analysis. My intention was to comprehensively compare the performance across all components, ensuring no critical information was overlooked in the initial stages of model development.

3.2 Hyperparameter tuning

I focused on fine-tuning two models: Random Forest and Gradient Boosting, with the target variable being Total_Medal_Score and the remaining variables serving as predictors. Prior to the tuning process, I conducted a z-score test to evaluate the statistical significance of each feature in relation to Total_Medal_Score. The results of this test indicated that most features except NOC factors had a z-score less than 0.1, demonstrating their statistical significance.

For the Random Forest model, my primary tuning parameter was the number of trees. I varied the number of trees from 50 to 1000, incrementing in steps of 50. This process was repeated nine times, as I iteratively dropped variables in each trial to assess their importance and determine whether any features could be excluded. The outcomes of these trials revealed that the model achieved the smallest Mean Squared Error (MSE) when all predictors were used, indicating that each contributed valuable information to the model.

In the case of the Gradient Boosting model, I employed cross-validation with 5 folds to ensure the robustness of my findings. The hyperparameters tuned included the number of trees (ranging from 50 to 400, in steps of 50), the depth of the trees (ranging from 1 to 5, in steps of 1), the learning rate (tested at 0.001, 0.01, 0.05, and 0.1), and the minimum number of observations in a node (tested at 5, 10, and 15). Given Gradient Boosting's sensitivity to the scale of input data, I used the scaled dataset for fitting this model. This approach allowed me to optimize the model's performance, adjusting various parameters to find the most effective combination for predicting the Total_Medal_Score.

3.3 Model selection using R^2

I utilized R-squared as the evaluation metric to assess the performance of the models. This metric was chosen because it can provide a clear measure of the model's explanatory power.

For the Random Forest model, the best performing variant was the one that utilized all available predictors with the number of trees set at 350. This model configuration resulted in an R-squared value of 0.8511, indicating a high level of predictive accuracy. To further explore the model's capabilities, I applied the same hyperparameters to fit the scaled dataset and the dataset reduced by PCA. Interestingly, while the scaled dataset did not improve the R-squared (achieving 0.8469), the PCA dataset with 12 components significantly enhanced the model's performance, reaching an R-squared of 0.9576. Even with a reduced component count of 6 in the PCA dataset, the model maintained a high R-squared value of 0.9329, demonstrating the effectiveness of dimensionality reduction in this context.

In contrast, the best-performing Gradient Boosting model, characterized by 200 trees, a depth of 5, a learning rate of 0.1, and a minimum of 5 observations in a node, achieved an R-squared of 0.8047. While this was a respectable outcome, it was not as high as the best Random Forest models.

Consequently, the optimal model for my project, based on the R-squared metric, is the Random Forest model applied to the PCA dataset with 12 components, which achieved an R-squared of 0.9576. This model not only offered the highest explanatory power but also

indicated the effectiveness of combining Random Forest with PCA for predictive accuracy in this specific Olympic dataset analysis.

4 Results

4.1 Predictive Performance

The performance of the selected Random Forest model, particularly when applied to the PCA-reduced dataset, offers insightful revelations about the dynamics of Olympic medal predictions. Achieving an R-squared of 0.9576 with the 12-component PCA dataset signifies that the model can explain approximately 95.76% of the variability in the Total_Medal_Score, which is a substantial portion. This high level of accuracy indicates that the predictors, once reduced and optimized through PCA, are highly effective in forecasting Olympic outcomes. The reduction to 12 components suggests that a distilled set of features captures the essential information needed for accurate predictions, highlighting the power of dimensionality reduction in enhancing model performance.

The Gradient Boosting model, while showing respectable performance with an R-squared of 0.8047, fell short of the Random Forest model's accuracy. This difference could be attributed to the intrinsic strengths of the Random Forest algorithm in handling this specific type of dataset, which might be better suited to capturing the complex, multi-faceted nature of Olympic performance data.

In summary, the high R-squared value achieved by the Random Forest model with PCA indicates a strong predictive power, suggesting that the model is highly effective in understanding and forecasting Olympic medal outcomes.

4.2 Significance of Predictors

Feature importance in this context is measured by two metrics: the percentage increase in Mean Squared Error (%IncMSE) and the increase in Node Purity (IncNodePurity) when each feature is used for splitting in the Random Forest algorithm.

- **NOC:** With a relatively low %IncMSE of about 6.04%, the National Olympic Committees (NOC) feature seems to have a modest impact on the model's predictive performance. Its IncNodePurity is also quite low, suggesting it doesn't contribute much to homogenizing the data splits in the model.

- **Comp.1:** This component stands out with the highest %IncMSE of approximately 37.70%, indicating it is the most significant predictor in the model. Its high IncNodePurity also corroborates this finding, showing that it greatly contributes to the purity of the model's decisions.

- **Comp.6:** The sixth component is also notable, with a %IncMSE of about 24.12%, making it the second most important feature for the model's predictions. Its IncNodePurity is the highest among all components, which implies that it is particularly effective at creating pure splits.

- **Comp.7:** This component has a moderate level of importance with a %IncMSE of around 14.94% and a reasonably high IncNodePurity, suggesting a solid contribution to the model's performance.
- **Comp.4 and Comp.5:** These components have a substantial impact on model performance, with %IncMSE values around 10.16% and 12.40%, respectively. They also contribute to node purity, indicating their usefulness in the model.
- **Comp.2, Comp.3, Comp.8, and Comp.9:** These components have lower %IncMSE values ranging from about 4.57% to 9.97%, suggesting they have less influence on the model's prediction accuracy compared to the most important components.
- **Comp.10 and Comp.11:** These components have the least impact on the model, with very low %IncMSE values around 1.97% and 7.94%, respectively. Their contributions to node purity are also on the lower side.

5 Conclusion

This project set out with the dual objectives of assessing the impact of the host city on the performance of athletes from the same country and developing a predictive model for future Olympic medal outcomes at the country level. Through rigorous data preparation, which involved careful cleaning and strategic feature engineering, a refined dataset was produced, enabling a more focused analysis of the factors at play.

The exploration and preprocessing stages revealed the significance of the dataset's features and provided a foundation for model development. By utilizing Random Forest and Gradient Boosting models and fine-tuning their hyperparameters, it was determined that Random Forest yielded superior performance, particularly when paired with Principal Component Analysis for dimensionality reduction.

The optimal model, a PCA-reduced Random Forest, demonstrated an impressive ability to explain about 95.76% of the variability in the Total_Medal_Score with its highest R-squared value. This model's strength lies in its ability to distill complex relationships within the data into an accurate predictive tool. Its robustness was further highlighted when compared to other model configurations and to the Gradient Boosting model.

The insights gained from this model not only shed light on the historical performance of countries in the Olympics but also provide a reliable method for forecasting future outcomes. The high degree of accuracy suggests that the model can be a valuable asset for national sports committees and Olympic teams in strategizing and allocating resources efficiently.

In conclusion, this project has successfully developed a predictive model that captures the multifaceted nature of Olympic performance and provides a meaningful tool for anticipating future trends. As new data emerges from subsequent Olympic Games, this model can be updated and potentially improved, ensuring its relevance and utility in an ever-evolving sporting landscape. The methodologies and findings of this study contribute to the broader field of sports analytics, highlighting the potential of machine learning techniques in enhancing our understanding of athletic performance on the world stage.

6 Appendices

6.1 Figure

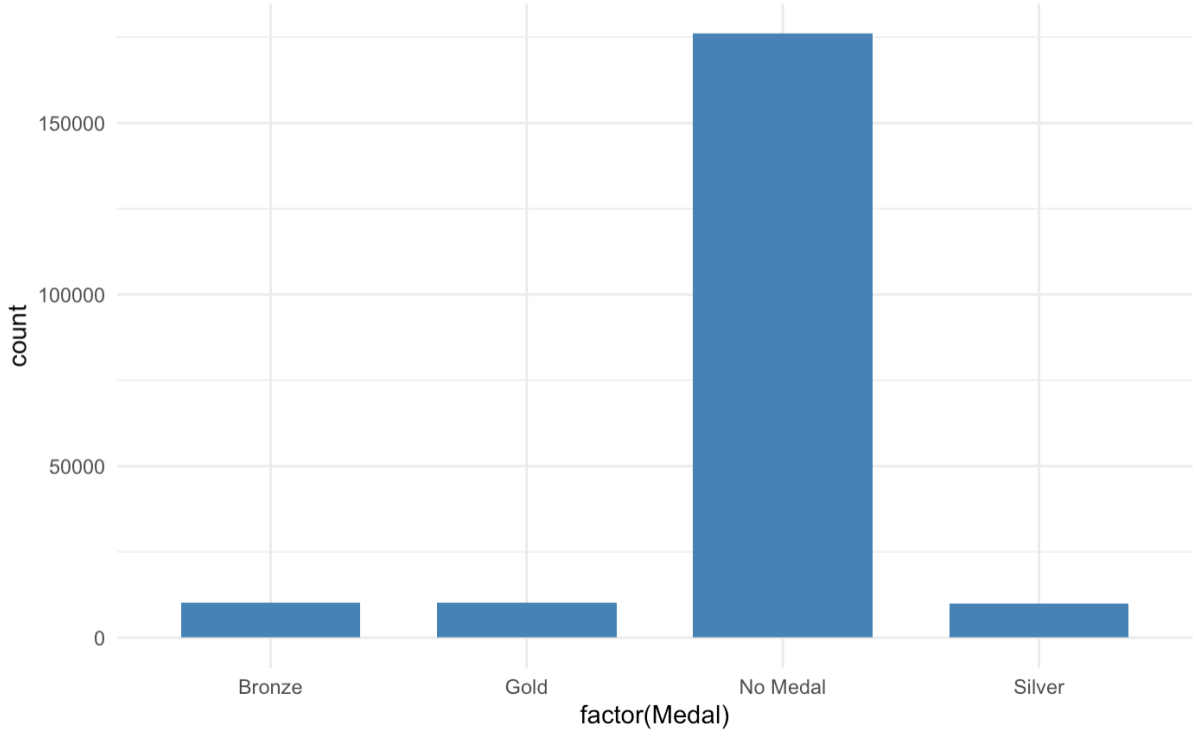

Figure 1: Bar Plot of Medal counts

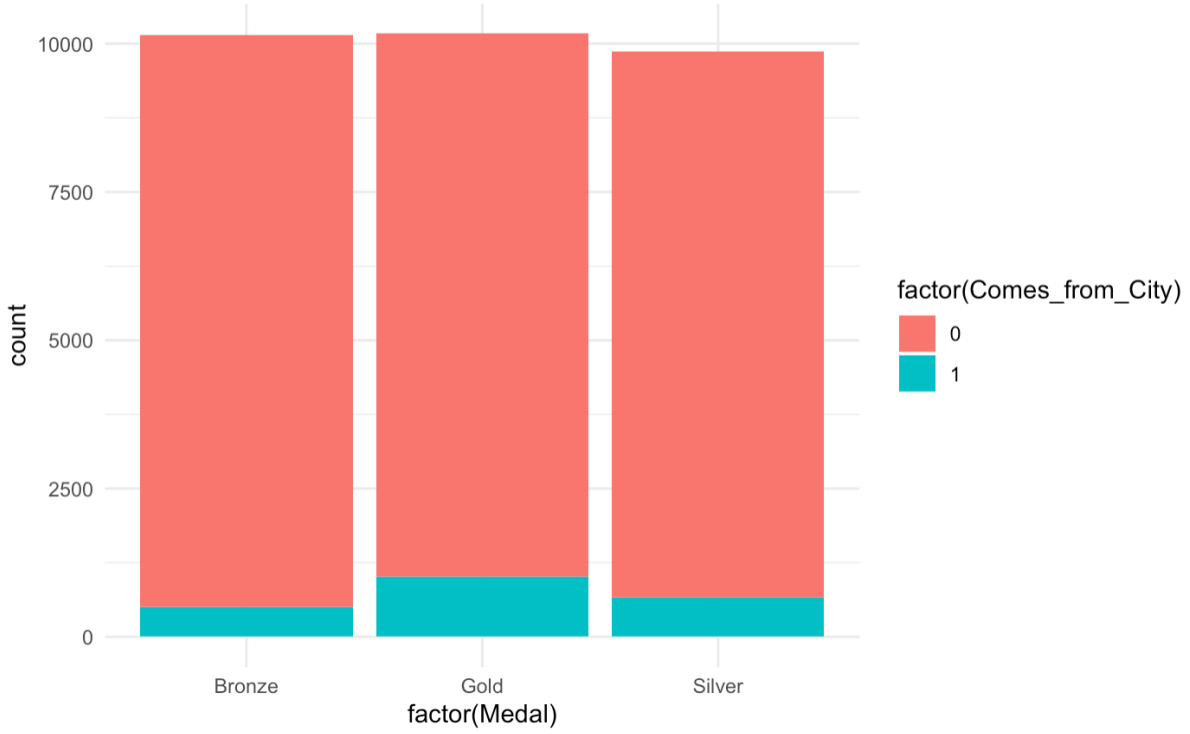

Figure 2: Stacked Bar Plot of Medal counts

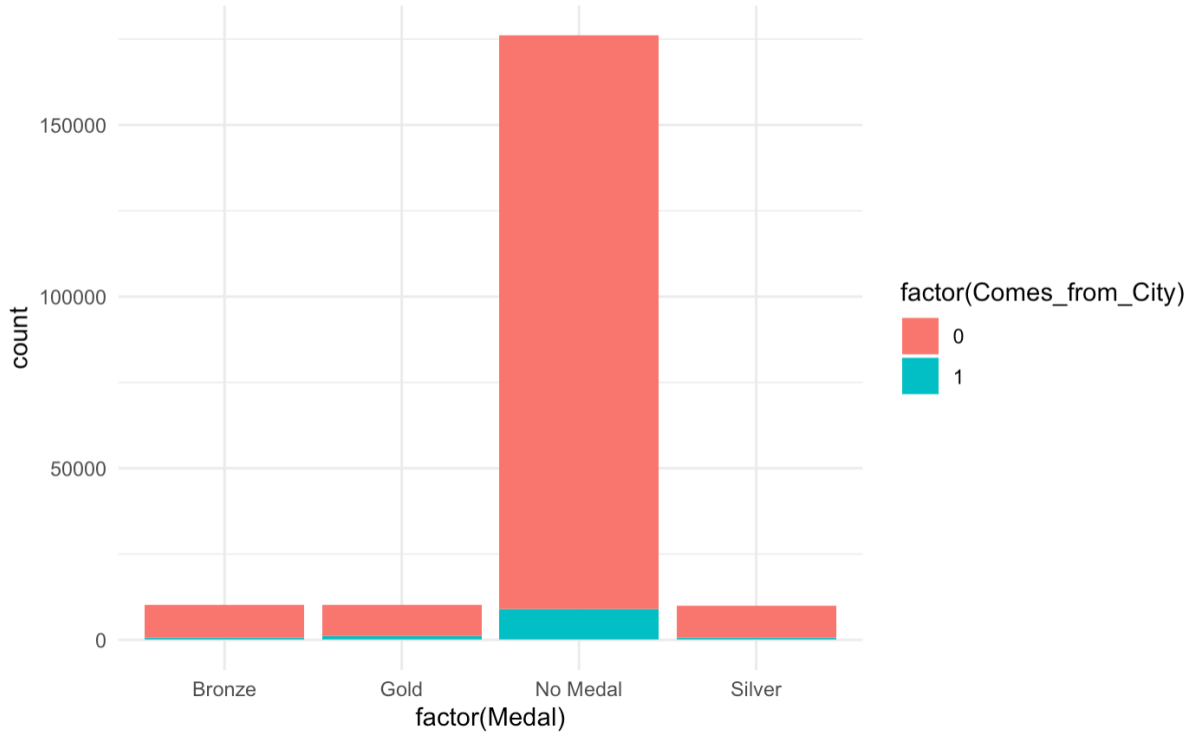

Figure 3: Stacked Bar Plot of All Medal counts

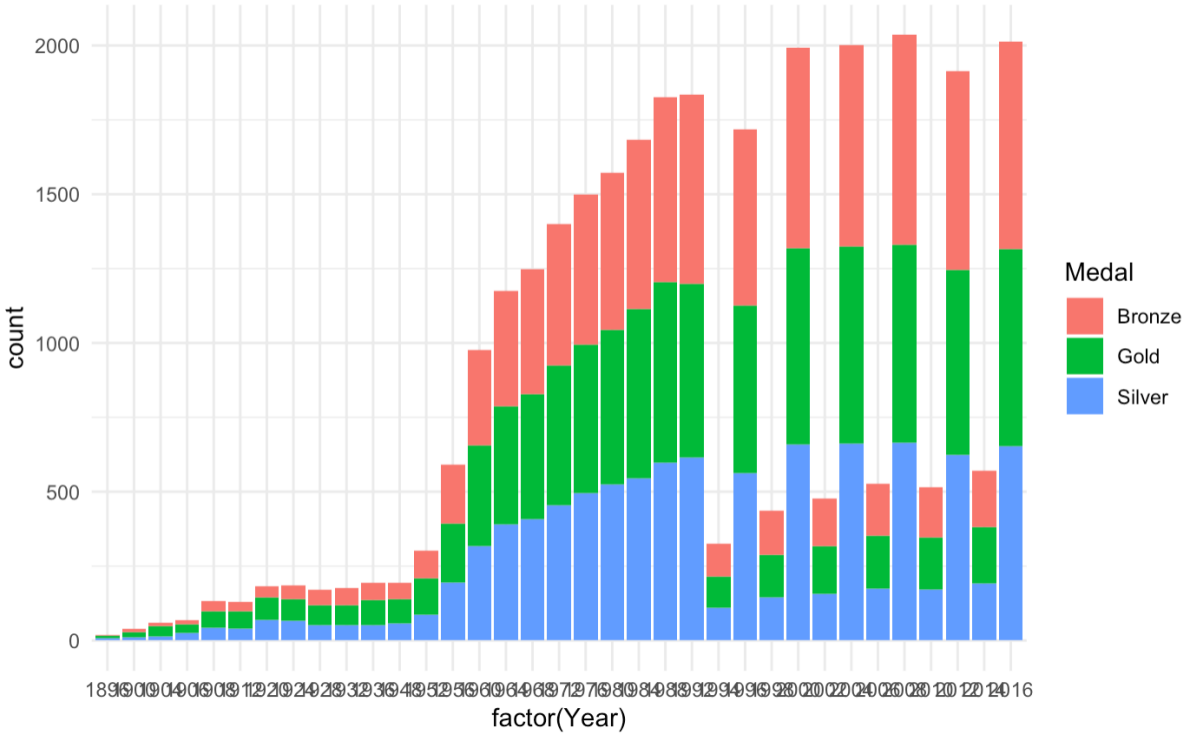

Figure 4: Stacked Bar Plot for yearly medal counts

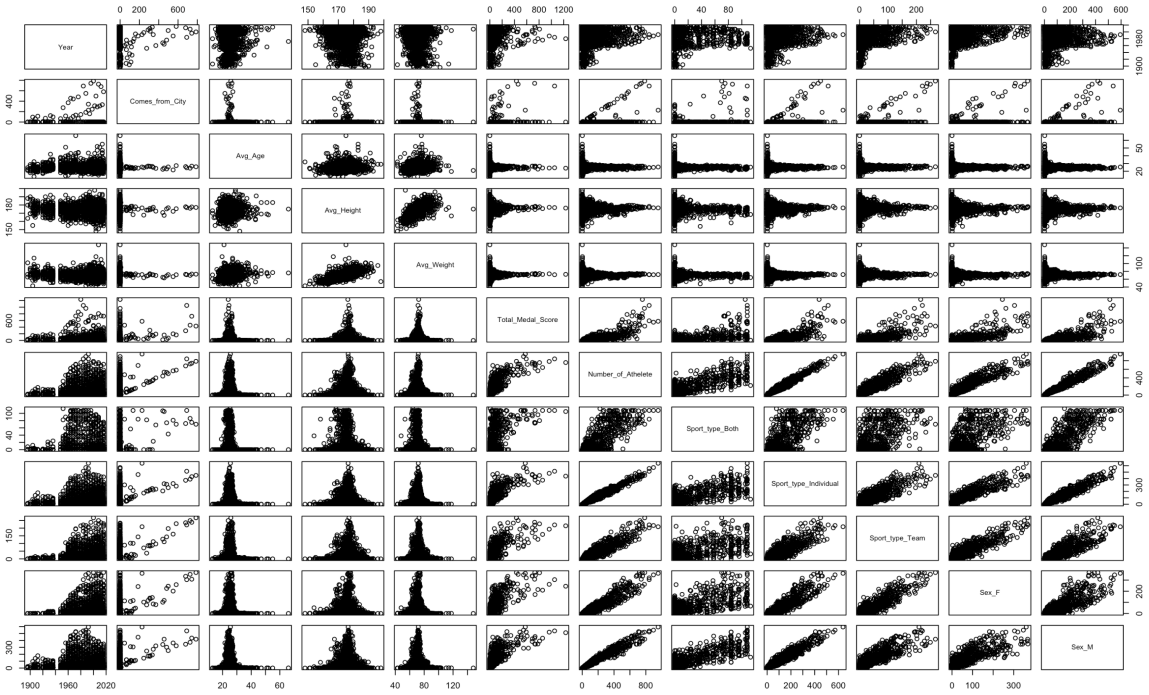

Figure 5: Pair plot of all the predictors

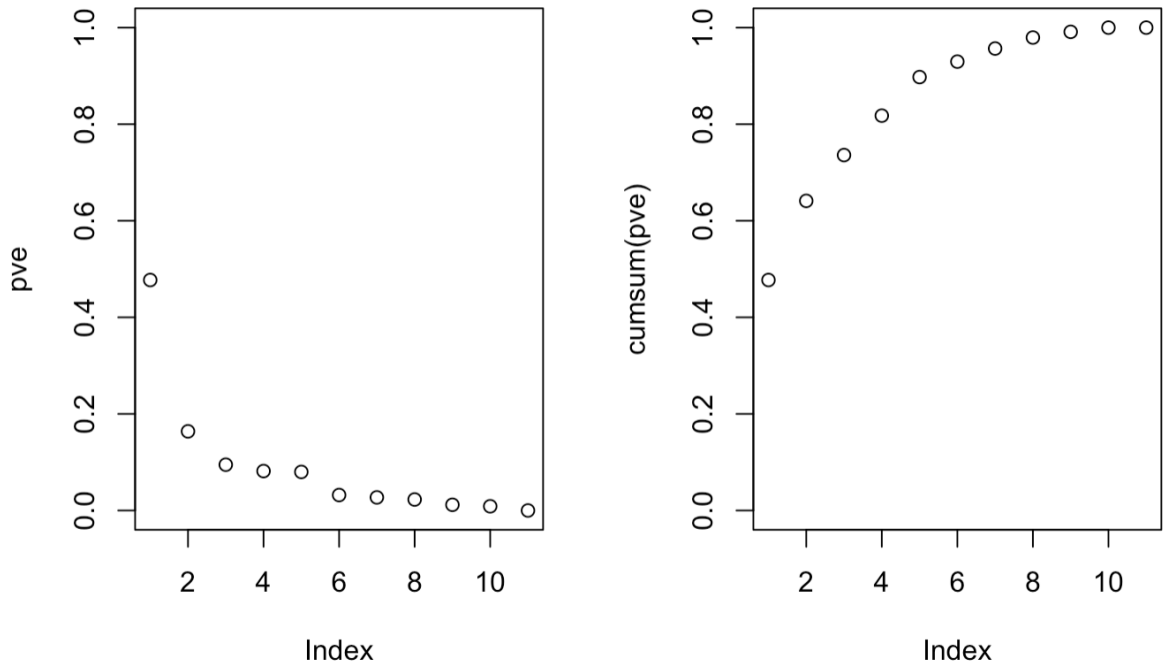

Figure 6: Scree Plot for PCA