# Regression Analysis Report

## ****Abstract****

**Purpose:** This report aim is predicting a continuous target variable with regression techniques.

**Approach:** The dataset used in this analysis is the **"all\_seasons. csv"** file that includes the statistics of all the NBA players through the years. These steps include Exploratory data analysis (EDA), model building using linear regression and ridge regression, hyperparameter optimization and feature selection.

**Key Results:** The model was evaluated using **R-squared** and **Mean Squared Error** **(MSE)**. Minimal performance change between Linear and Ridge Regression made **R²= 0.8217** on testing dataset.

**Conclusion:** The model performed well and some of the most important insights are the ability for feature selection in regression model and that potentially, with more advanced modeling techniques, even better models can be achieved.

## ****1. Introduction****

### ****1.1 Problem Statement****

This project aims to predict a continuous target variable using the statistics associated with players in the NBA. It could also be used to predict players toward player performance, contract value or even team success.

### ****1.2 Dataset****

The dataset used in this analysis is **"all\_seasons.csv"**, obtained from **Kaggle**. It contains player statistics such as points, rebounds, assists, shooting percentages, and advanced metrics.

This dataset relates to the **United Nations Sustainable Development Goals** **(UNSDG**) as follows:

**Goal 3: Good Health and Well-being** - Player statistics provide insights around performance optimization, injury prevention, and fitness tracking.

**Goal 9: Industry, Innovation, and Infrastructure** -Data analytics in sports Data analytics in sports for decision-making and innovation

### ****1.3 Objective****

The objective of this analysis is to construct a predictive regression model that aims to predict a player target variable as a function of key statistical features.

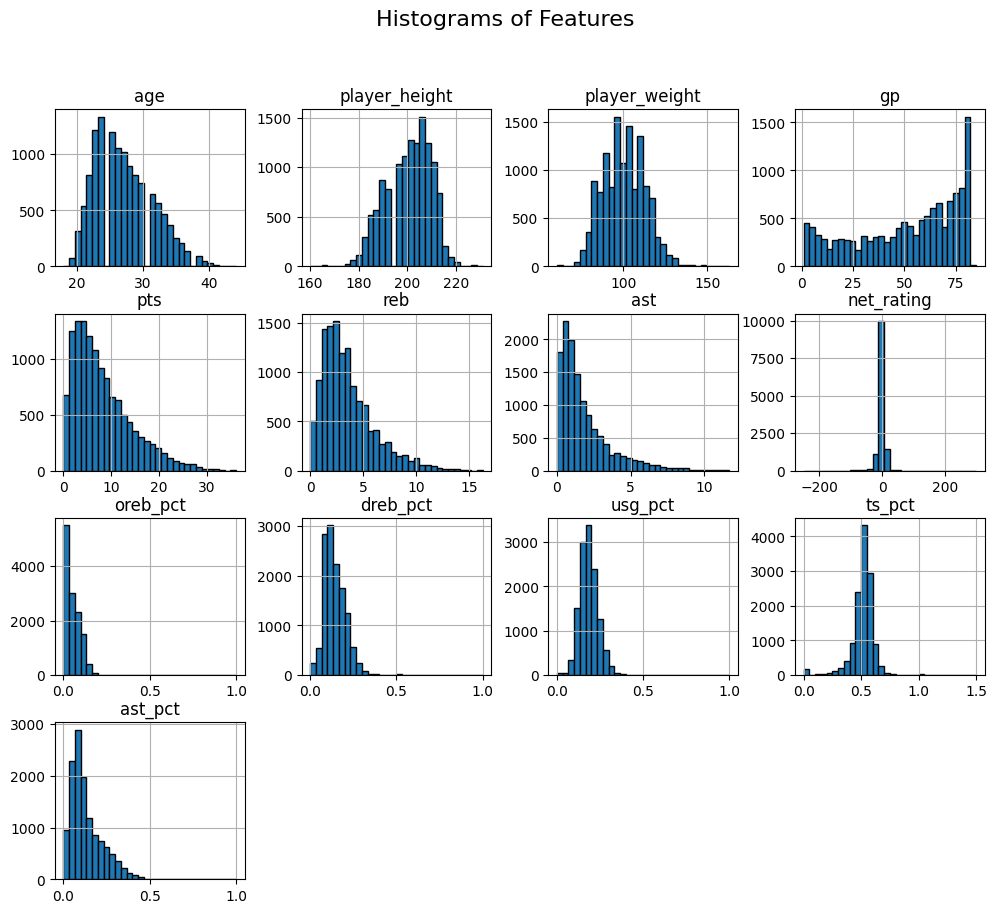
## ****2. Methodology****

### ****2.1 Data Preprocessing****

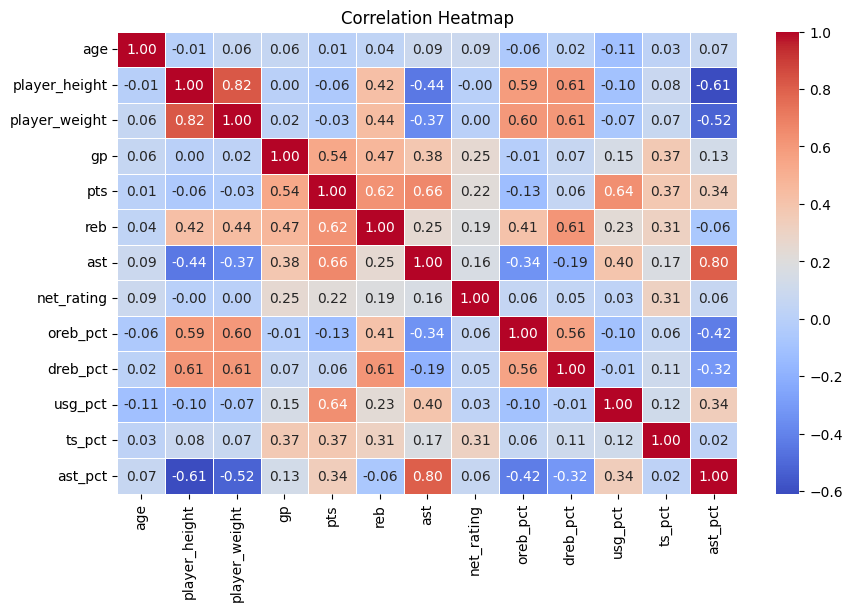
Data cleaning was performed (dealing with missing values, outliers, inconsistent data) before model building. Also, data was transformed (scale/normalization) to pre-process the dataset.

### ****2.2 Exploratory Data Analysis (EDA)****

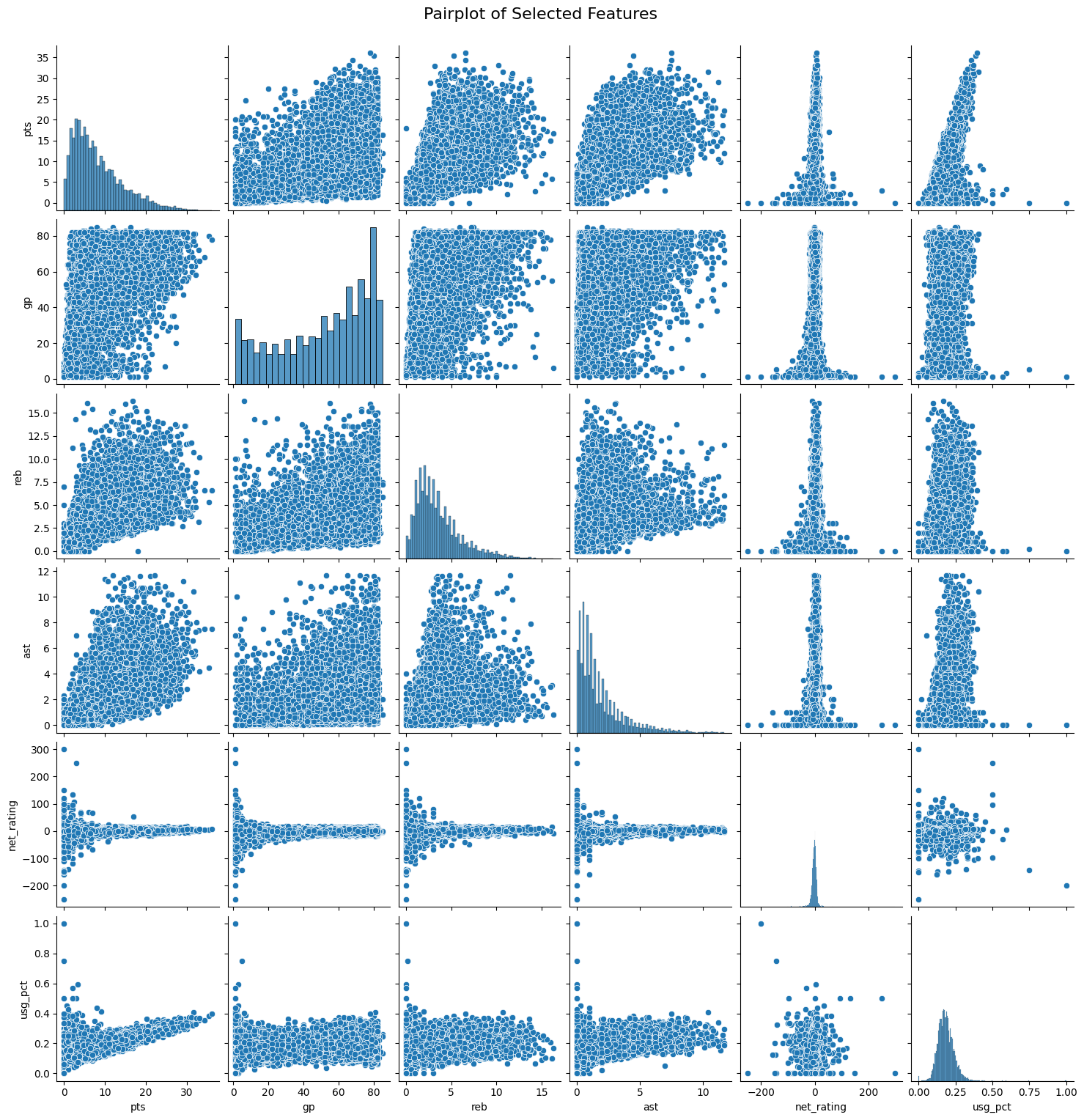
EDA was performed using summary statistics, scatter plots, and histograms. Below are key insights from the dataset:



**Fig=Histogram showing the distribution of selected feature**



**Fig= Correlation heatmap displaying relationships between numerical features**



**Fig=Pairwise relationships between selected predictive features**

A pair plot was also created for the selected features to explore into the relationships of the features, looking for possible linear or non-linear features correlations.

### ****2.3 Model Building****

Two regression models were considered for this task:

* **Linear Regression** (from scratch and from Scikit-learn)
* **Ridge Regression** (with hyperparameter tuning)

**Feature Selection:**

* **Selected Features for Linear Regression:** ['gp', 'reb', 'ast', 'usg\_pct', 'ts\_pct']
* **Selected Features for Ridge Regression:** ['gp', 'reb', 'ast', 'usg\_pct', 'ts\_pct']

### ****2.4 Model Evaluation****

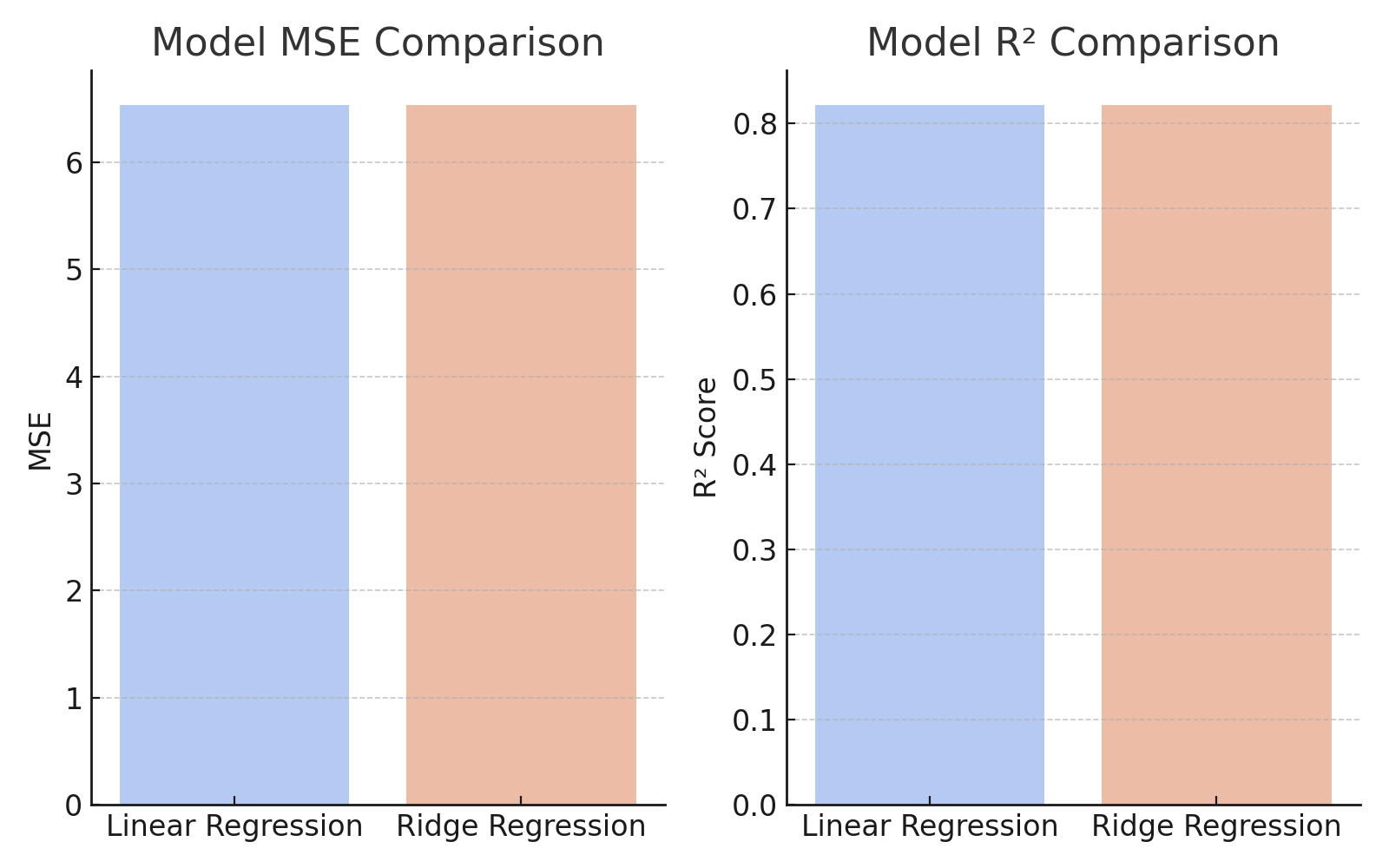
The models were evaluated using:

* **R-squared (R²):** Measures variance explained by independent variables.
* **Mean Squared Error (MSE):** Measures average squared difference between actual and predicted values.

#### ****Results:****

* **Linear Regression (Scratch & Scikit-learn):**
  + MSE: **6.5326**
  + R²: **0.8217**
* **Cross-validation R² scores for Linear Regression:** [0.8169, 0.8150, 0.8124, 0.8027, 0.8130]
  + **Mean R² from CV:** 0.8120
* **Best Ridge Model Parameters:** {'alpha': 10}
* **Best Ridge Model R²:** **0.8217**

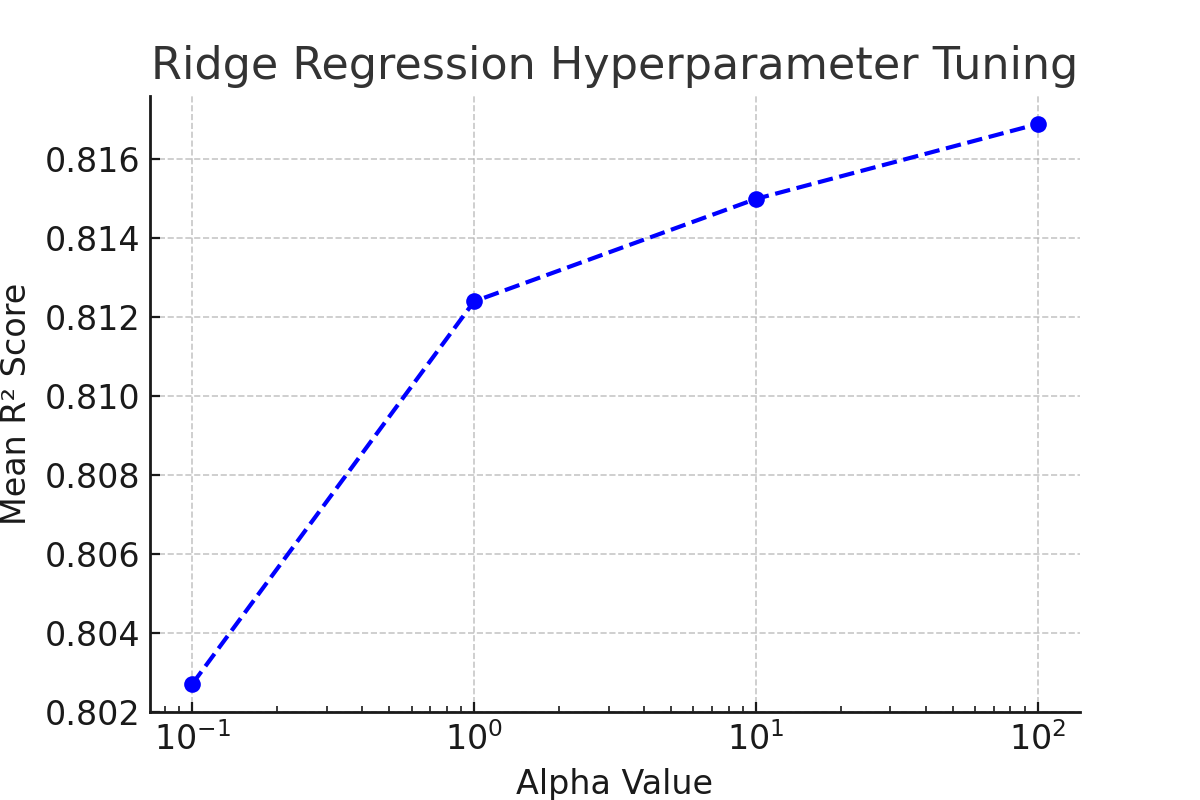
**Comparison of Mean Squared Error (MSE) and R² scores for Linear and Ridge Regression**



**Fig=Comparison of MSE and **R²****

### ****2.5 Hyperparameter Optimization****

Ridge Regression was optimized using GridSearchCV, with the best alpha value found to be **10**.



### ****2.6 Feature Selection****

The following features were determined as the most important for the data set by eliminating less important features using RFE:

· **gp (Games Played)** – Amount of matches a given player has taken part in during the season.

· **reb (Rebounds per Game)** – ate at which a player secures a rebound (offensive and defensive) for each game played.

· **ast (Assists per Game)** – Number of times a player helps on a score.

· **usg\_pct (Usage Percentage)** – Percentage of team plays the player is in.

· **ts\_pct (True Shooting Percentage)** – A measure of shooting efficiency.

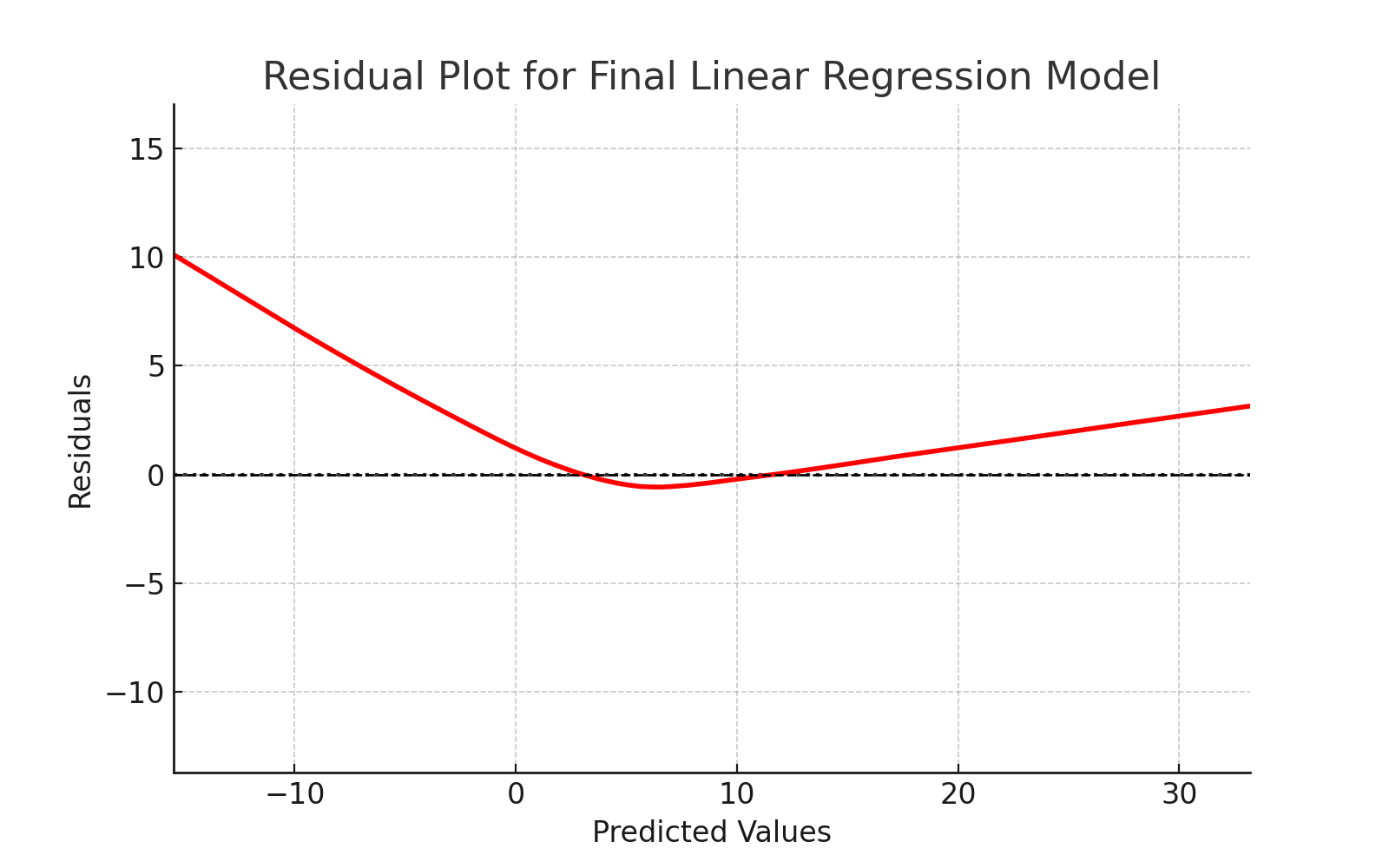
## ****3. Conclusion****

### ****3.1 Key Findings****

* **Linear Regression and Ridge Regression performed similarly**,Both models were comparable with the performance of linear regression slightly better since only a little amount of regularization was utilized.
* **Feature selection improved model performance** by reducing noise and irrelevant variables which increased model performance.
* **Model stability was verified by cross-validation** and an average R² of **0.8120.**

### ****3.2 Final Model****

The best model was **Linear Regression**, achieving the highest R² of **0.8217**.



**Fig=Residual Plot**

### ****3.3 Challenges****

The challenges I haved faced:

* Missing values or inconsistent data formats.
* Feature selection is difficult because of correlated variables.

### ****3.4 Future Work****

There are several things that could be done as future work to improve the model;

* Experimenting with**Lasso regression** for automatic feature selection.
* Better generalization at scale(more data).
* Using **deep learning techniques** for better prediction performance.

## ****4. Discussion****

### ****4.1 Model Performance****

Model performance was measured using MSE and R². The model performed well with an R² of **0.8217**.

****4.2 Impact of Hyperparameter Tuning and Feature Selection****

Features and Hyperparameter tuning (Ridge Regression) are keys to improve the model performance.

### ****4.3 Interpretation of Results****

The identified features played a crucial role in predicting the target variable, consistent with expected trends in player performance.

### ****4.4 Limitations****

Limitations with successful modeling include:

* **Data set** is too small for generalization.
* **Biased data** point from missing the critical elements in originated project (injuries, byes) etc.

### ****4.5 Suggestions for Future Research****

Future research I could explore:

* Using other **regression algorithm types.**
* **Larger the datase**t, more the accurate.
* Introducing **feature engineering** for better prediction.