DASC 5133 INTRO TO DATA SCIENCE

Project Report of Data Analysis of Fantasy Premier League

Instructor	Dr. Yalong Wu
Student Name	Karthik Jallepalli Bhanu Shankar Kantati Yashwanth Manglarapu Praneeth Goud Poshala Chetan Sylaka
Student Email	Jallepallik0700@uhcl.edu Kantatib0227@uhcl.edu Manglarapuy0691@uhcl.edu Poshalap0725@uhcl.edu Sylakac7446@uhcl.edu

Abstract: This report aims to describe about the Design and implementation of the Fantasy Premier League analysis. Here we work towards to achieve the prediction of the players who are to score high in the future game week by analysing the historical data and statistics. It is to extract the best players to consider for purchase and predict the total points of the game.

Keywords: FPL – Fantasy Premier League

1. Introduction

1.1 Project Introduction

Football, which is one of the sports games loved by most of the people all over the world. It has changed substantially over time, both in terms of its on-field dynamism and its virtual manifestation. The well-known fantasy football game Fantasy Premier League (FPL) gives fans a special chance to interact with the game in a data-driven, statistical, and strategic way. To win the FPL championship, users, or users or customers must build their ideal teams, manage player rosters, and outsmart opponents. We explored the data that was related to the English Premier League season, which runs from August to May each year.

FPL enthusiasts construct virtual teams with real-world footballers, and the team wins the points or loses the points based on their performance on the real-time field. By using this analysis, we can help people to make data driven decisions and understand the game stats better and they can easily predict the player who will score more goals and the condition and form of the player by the strategic by attaining required information from then analysis report. Additionally, it can also be used for predictive analysis for further improvements.

Here we analyse the existing historical game data by executing required data science life cycle tasks and ultimately, build a Machine Learning model that can predict the points of the players in upcoming game week.

1.2 Objectives

The purpose of this project is to derive insights from Fantasy Premier League's previous data and aid participants (users) in picking players for the league based on data. Here, we present the more complicated facts along with simply the information that is necessary clearly to everyone. To better assess the player's performance and ensure that every dollar they are paying is worthwhile, we visualize the results using graphs and charts to understand patterns and trends in any data, for example, price is very volatile

Tasks to do

- Cleaning the data to understand it better.

- Exploratory Data Analysis.
- To achieve findings of correlation among the attributes.
- Visualization of the Data and Results.

People can benefit from this analysis report by better-comprehending player statistics and gaining confidence when choosing players and formulating plans. The analysis of the players' strengths and limitations and how to compensate for them will be greatly aided by this knowledge. They can make the best transfer strategy decisions with all the facts at their disposal. This report is beneficial to a vibrant FPL community where everyone can express their thoughts and engage in dialogue.

1.3 Usage of This Report

This report delivers the complete understanding of the Data Science Life Cycle concept from primal to professional understanding. And how the data can be helpful to make an informed data driven decision.

2. Literature Survey

The main principle found in earlier publications has been to forecast future scores using historical statistical data and machine learning techniques. Every data analysis needs to all the basic steps to approach the goal i.e., Data collection and Data Preprocessing. To Understand the data very well and for modeling, we need to know about data behaviour, and that can lead to better modeling.

According to [1] The Gaussian Naive Bayes method and the statistics from prior game weeks were combined by Thapaliya to predict future performances, with a stated accuracy of 86 percent. He divided the data labels into two groups: those with six points or more and those with fewer than six. Here they know exactly what data needs to be extracted for modelling by exploratory data analysis.

[2] In Fantasy Premier League - Performance Prediction EDA was performed on 2019-20 season data in which the ROI (return on investment) of each position is plotted against ROI such as ROI-goal keeper, ROI Midfielders, defenders, and forwards. Only return on investment is taken into consideration. For prediction, only a 75% chance of playing is considered to avoid suspension, injured players, and unavailable players. some techniques used in this work to feature engineering are correlation matrix and threshold for each position.

[3] The optimal drafting problem in hockey pools, which is analogous to the drafting process in fantasy football, is examined in the 2007 study by Summers, Swartz, and Lockhart. The authors use a statistical method to calculate the likelihood that a player's lineup will win over another lineup at each stage of the draft. The best draft strategy is to select the available hockey player who increases this possibility. The player selection draft of a single real-world NFL franchise is the subject of a stochastic dynamic programming (DP) model proposed by Fry, Lundberg, and Ohlmann (2007).

In this model, the 3 best selection of drafting at each round is determined by the DP recursion that maximizes the sum of the value added by the drafted player and the total expected value added to the team in subsequent rounds. Some simplifying assumptions are added to the model

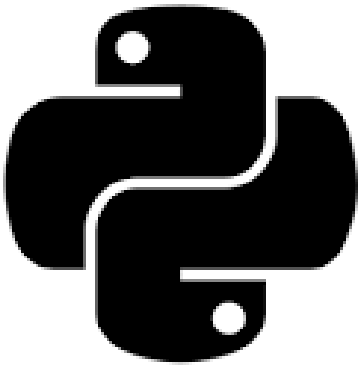
to decrease stochasticity (mostly the uncertainty in the conduct of the opponent teams) and shrink the state space to construct a computationally tractable model. The deterministic DP that results from this can be effectively solved as linear programs.

3. Software Requirements

Here the functional requirements are:

- Google Colab.

- Python 3.8



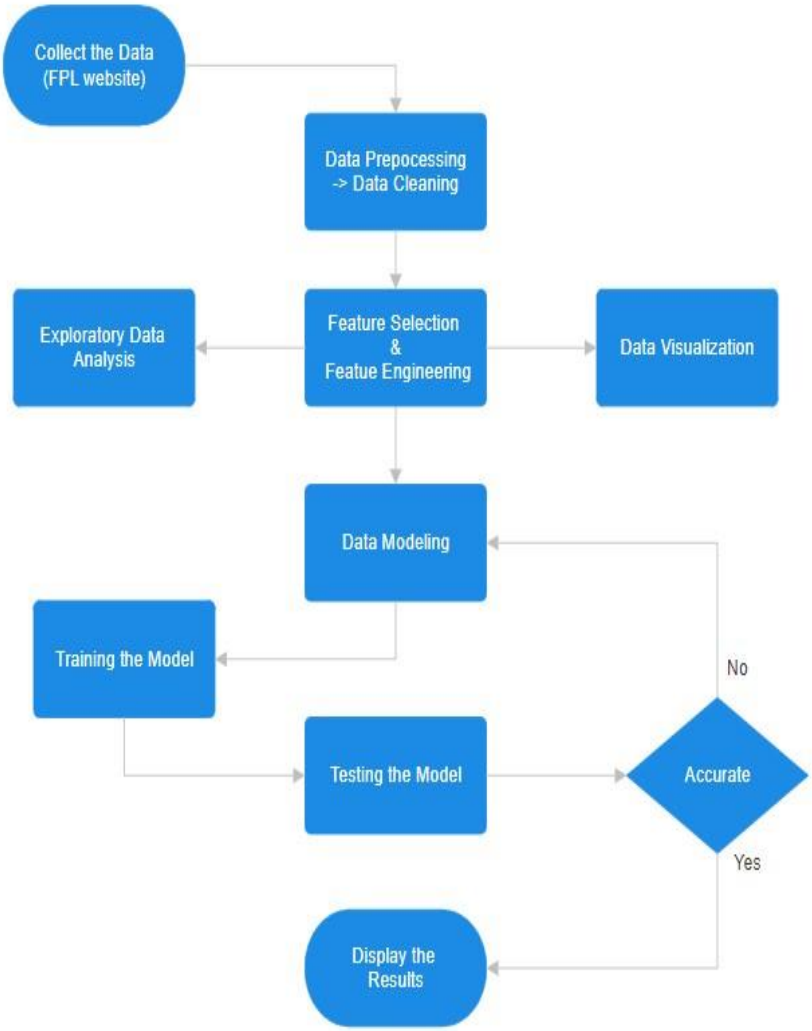
Libraries used:

- **pandas:** pandas are defined as an open-source library that provides high-performance data manipulation in Python. Pandas provides some sets of powerful tools like **DataFrame** and **Series** that are mainly used for analyzing the data, whereas in **NumPy** module offers a powerful object called **Array.**
- **numpy:** It is defined as a Python package used for performing the various numerical computations and processing of the multidimensional and single-dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.
- **seaborn:** seaborn is one of the amazing libraries for visualization of graphical statistical plotting in Python. Seaborn provides many color palettes and defaults beautiful styles to make the creation of many statistical plots in Python more attractive.
- **matplotlib:** The **matplotlib.pyplot** is the collection command style functions that make matplotlib feel like working with MATLAB. The pyplot functions are used to

make some changes to a figure such as creating a figure, creating a plotting area in a figure, plotting some lines in a plotting area, decorating the plot including labels, etc.

- **sklearn:** Scikit-learn (formerly scikits. learn and known as sklearn) is a free software machine learning library for the Python programming language. It is very widely used across all parts of the bank for classification, predictive analytics, and many other machine-learning tasks.

4. Design and Implementation



Our Project Flow and design Steps as follows:

A. Data Collection
B. Data Cleaning and Pre- Processing (Includes Feature Generation and Selection)
C. Exploratory Data Analysis
D. Data Visualization
E. Splitting the Data Set into Training and Testing
F. Data Modeling
G. Training and Testing the Data Model
H. Display the Results

A. Data Collection

Getting Data from Fantasy Premier League API:

To use the Fantasy Premier League API, send HTTP requests to their endpoints. To access FPL data, send GET requests to the appropriate endpoints, and we receive a JSON file in return.

```
import requests

# Define the URL of the FPL API endpoint for player data
url = "https://fantasy.premierleague.com/api/bootstrap-static/"

# Send an HTTP GET request to the FPL API
response = requests.get(url)
json_file = response.json()
type
```

From the FPL API "https://fantasy.premierleague.com/api/bootstrap-static/" a JSON file is extracted, keys of the json file are

{events, game_settings, phases, teams, total_players, elements, element_stats, element_types}

Element stats consists of meta data related to the elements data

Elements is the key data which defines the perfomance matrics of the player

Element type is the data related to the positions of the player

B. Data Cleaning and Pre-Processing

Pandas is powerful library for data cleansing, here JSON file we have received contained unnecessary data and noise, we have performed certain operations and methods to achieve our goal.

The following steps are performed in our analysis:

- Data Cleaning
- Data Transformation
- Dealing with Outliers
- Handling Categorical Attributes
- Data Integration
- Data Validation

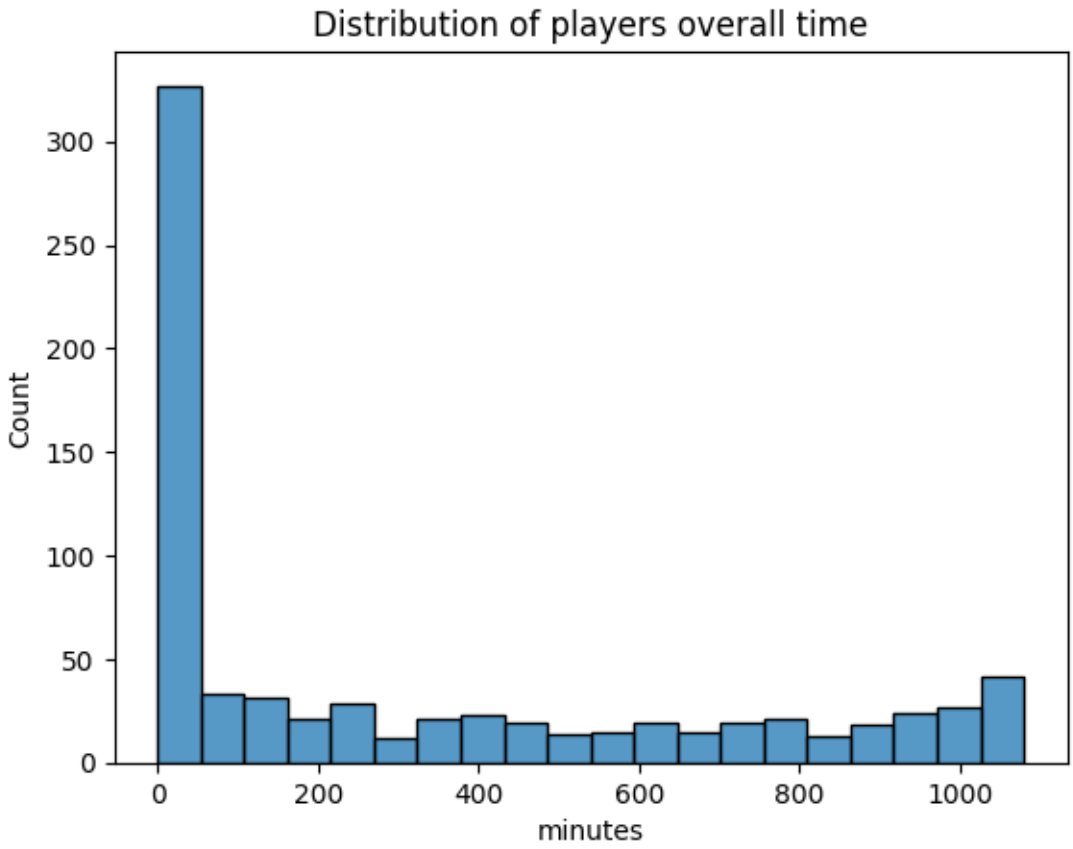
Obtained Elements data does not consist of the player description it has only the column of element type which is related to the element type data to get the detailed description of the player we need to merge both the elements and element type on element_type column.

Elements data consists of lot of columns from which we are considering the following columns [id,first_name, second_name, web_name , minutes , goals_scored , assists, now_cost, yellow_cards, red_cards, points_per_game, expected_goals, expected_assists, own_goals, element_type, points_per_game, total_points, plural_name_short, goals_conceded, penalties_saved,saves] for data analysis and modeling.

Most of the Attribute datatypes are object datatypes, which are unsuitable for analysis. For our convenience and usability of analysis function and operations we must convert eligible attributes into numeric datatypes.

C. Exploratory Data Analysis

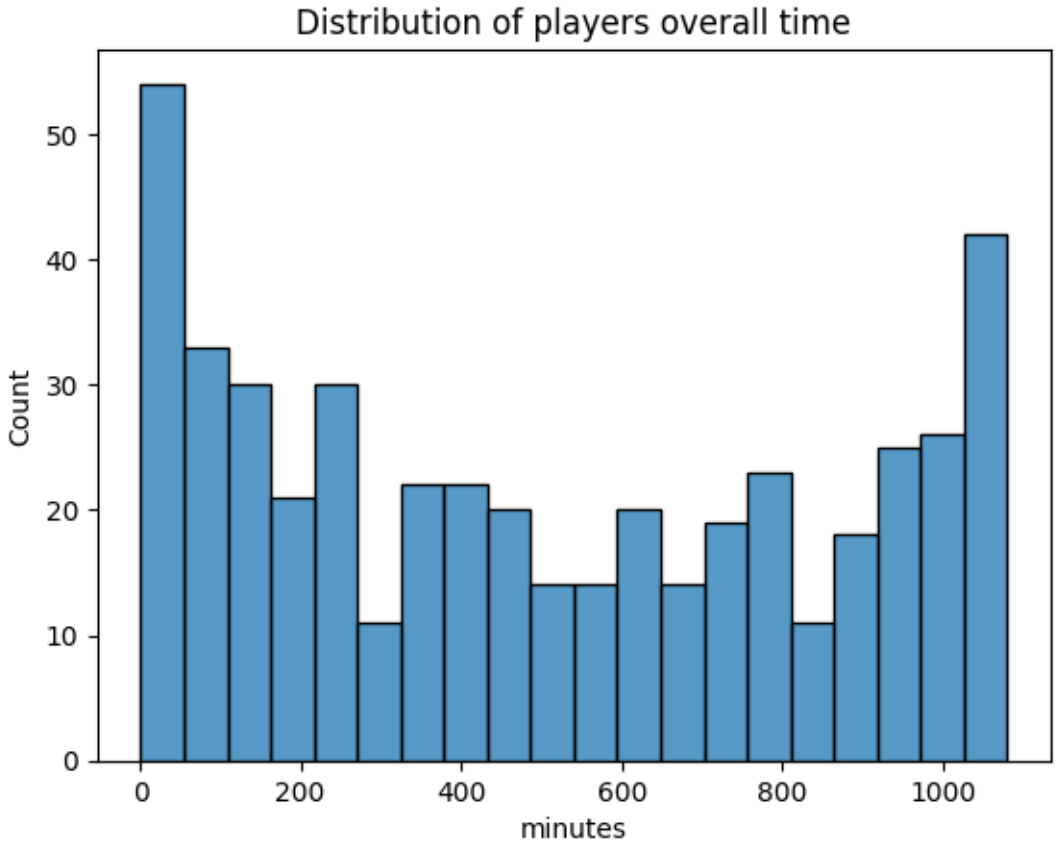
Exploratory Data Analysis is a phenomenon of exploring and understanding the Data parameters. We use data visualization or data descriptions for understandable display of the explored results.



Form the histogram mentioned above, we can see, it is clearly known that many of the player not even played more than zero minutes. It is essential to extract players with zero minutes for best results.

```
fpl_players = fpl_elements_copy[fpl_elements_copy['minutes']>0]
```

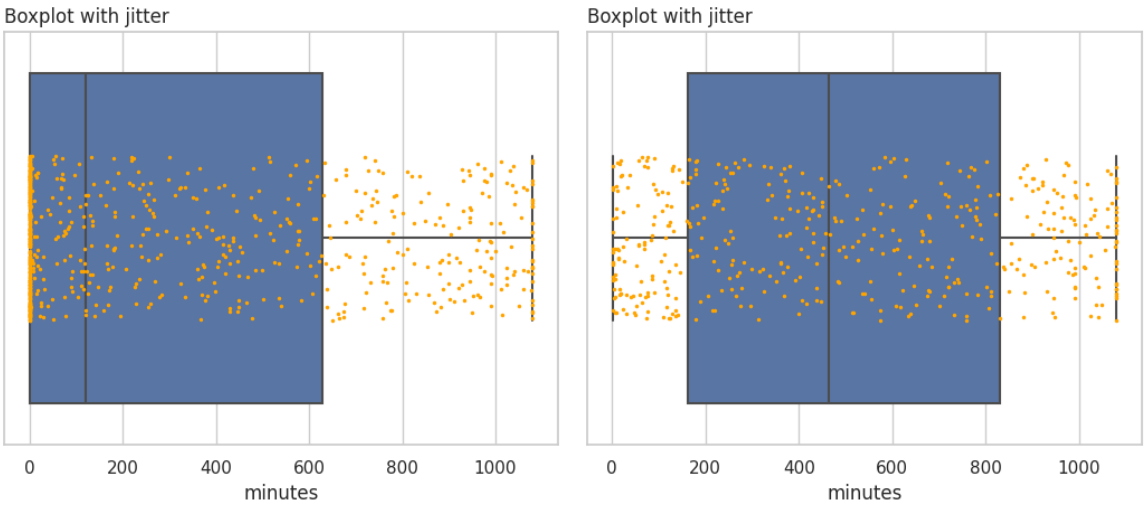
After the Data manipulation here are the results:



D. Data Visualization

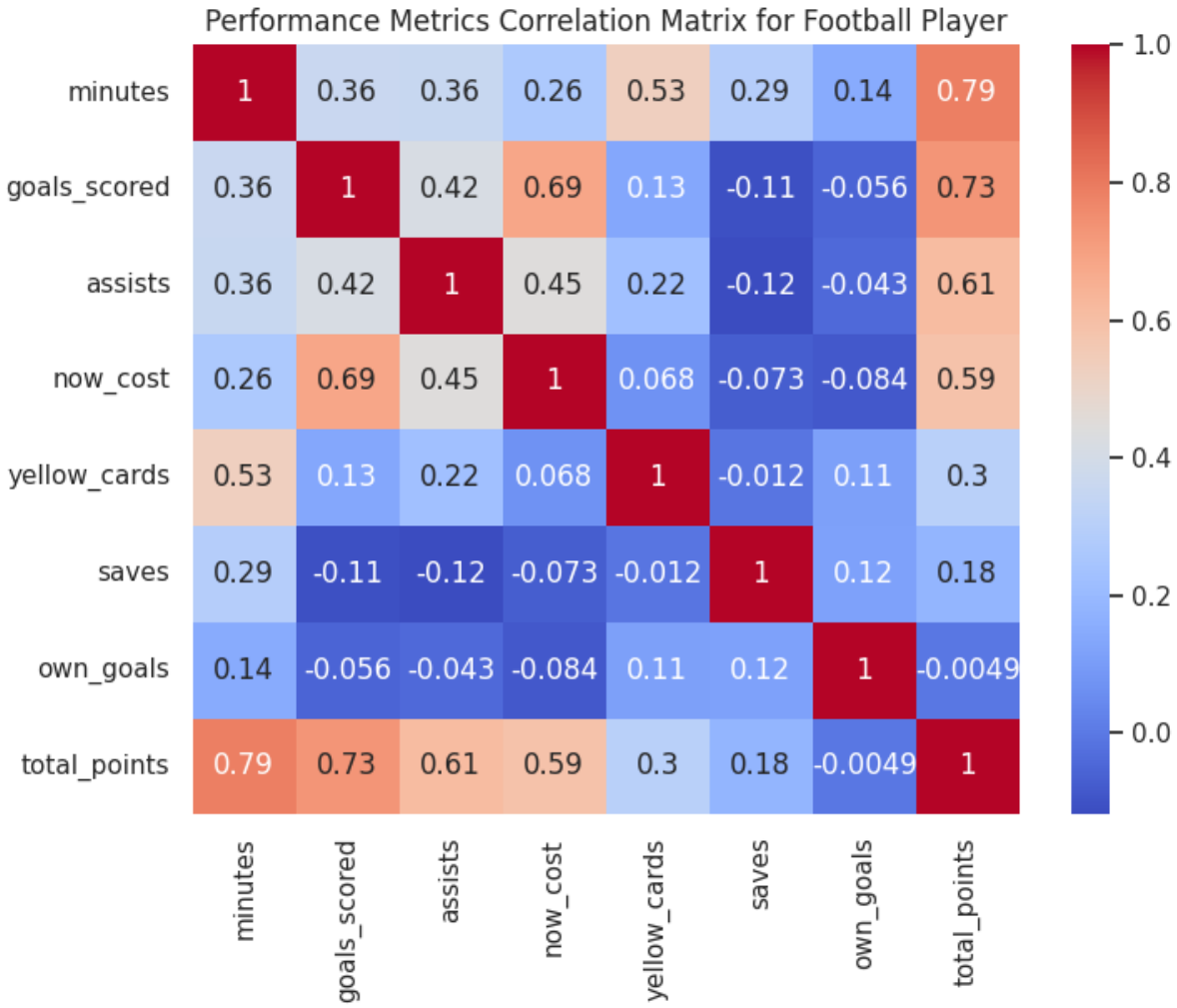
We can display the complex data by visualization to make it more understandable and visually appealing.

Distribution of Data Points and Nature of Data Before and After the Data Manipulation.



We can see in the above boxplots that skewness had changed from negative to the normal.

Using Seaborn library, we can find the corelation between the attributes which helps us understand the relations of the all the attributes and find the co-dependencies in the attributes.



Here Negative Value represents the indirect proposition among the attributes and Positive Values represent the Direct proposition of the attributes. As goals scored increased there is positive impact on the total points obtained. And also, we can see that saves and goal scored are negatively related.

From the correlation matrix it was known the total points has a correlation of more than 0.5 with the following attributes

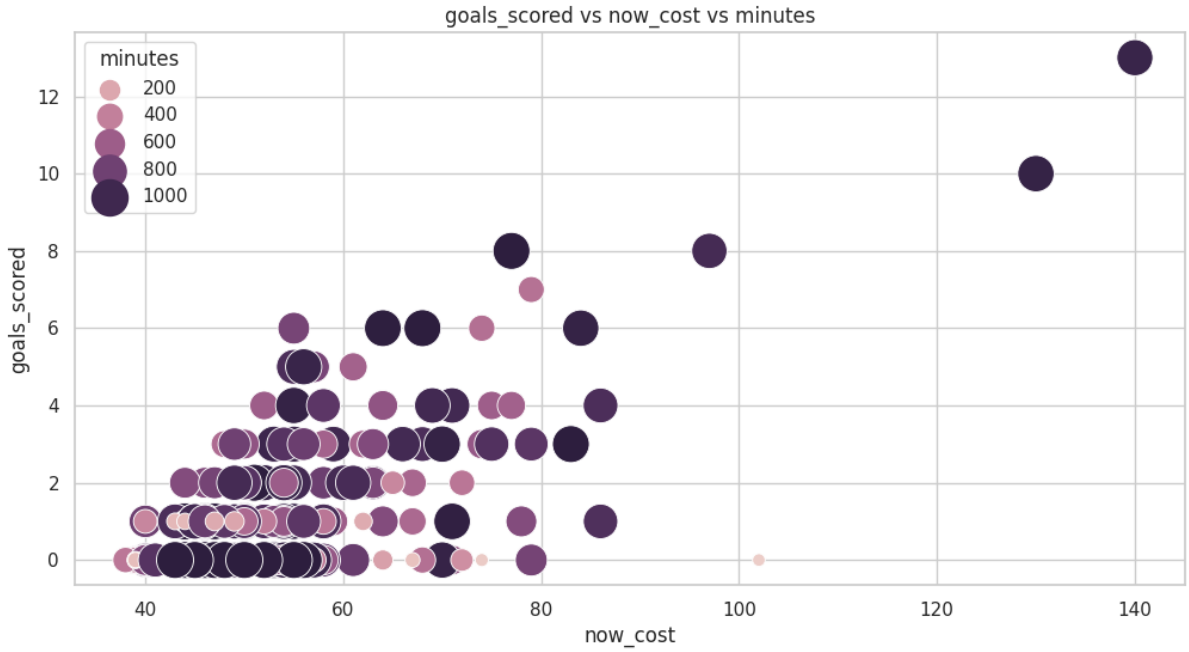
Number of minutes played = 0.79

Number of goals scored = 0.73

Number of assists = 0.61

Now cost = 0.56

Here we have used the seaborn library to display the spread of the data and parameter we used are goals_scored, now_cost and minutes played. Here minutes played can be displayed by the size of the bubble/point. We can see the relation among these three attributes below



E. Splitting the Data Set into Training and Testing

To implement the data into the data modeling we have to do the feature engineering and split the data into the training and testing data

For selecting the best players (MID, FWD, DEF)

- ✓ Players with a clean sheet (red cards = 0, 50th percentile of yellow cards) and the best performance are defined as
- ✓ Zero red cards
- ✓ less than the 50th percentile of yellow cards
- ✓ goals greater than the mean
- ✓ assists greater than the mean.

For acquiring undervalued players (FWD, MID, DEF)

- ✓ player costs less than mean selecting the best goalkeepers
- ✓ less than the 25th percentile of goals conceded
- ✓ saves more than the mean
- ✓ assists more than the mean

F. Data Modeling – Training and Testing of the Data Modeling

The Machine Learning algorithm we used here is Regression Analysis. Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

```
X = fpl_performance[['minutes',
'goals_scored','assists','now_cost','yellow_cards','saves','own_goals']
]
y = fpl_performance[['total_points']]
```

Independent variables of the data are minutes played, goals scored, assists, now cost, yellow cards, saves, own_goals Target variable is total points

```
# Import necessary libraries
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt


# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Create and train the model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)
```

```
# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

Here we have split the data set into 80% as Training and 20% as Testing Dataset.

For Model Validation:

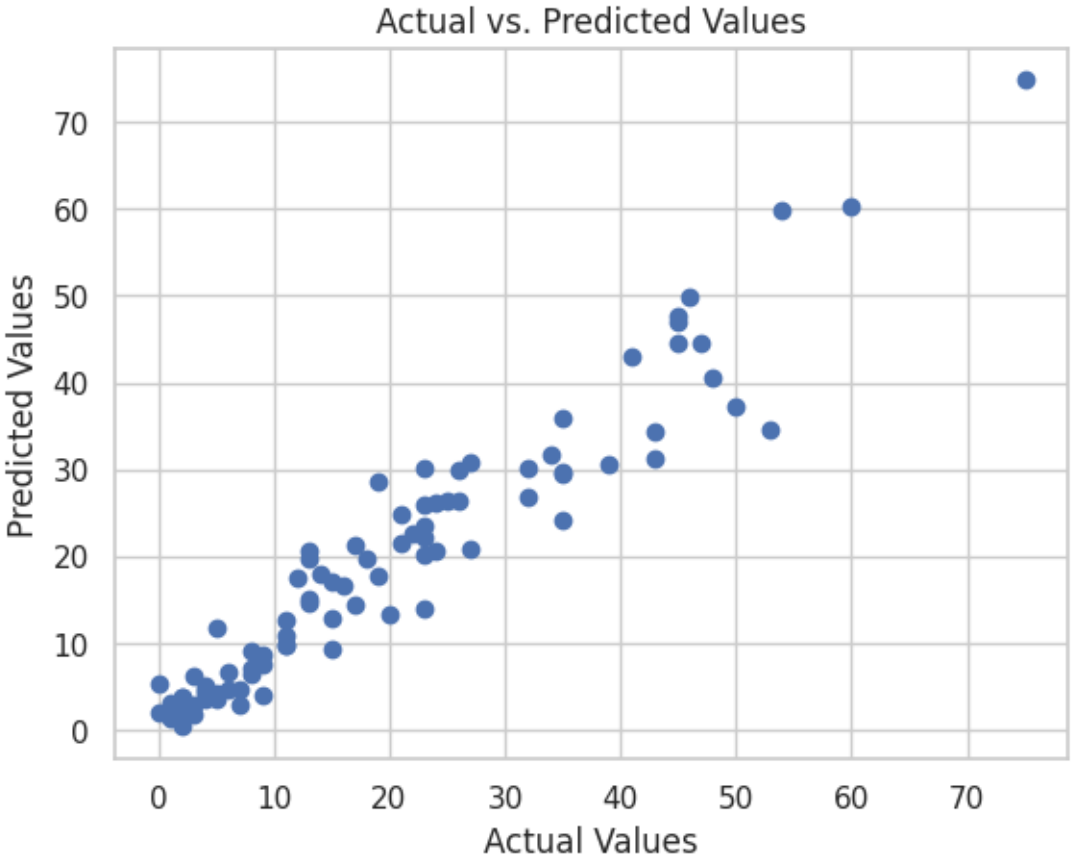
```
printf'Mean Squared Error: {mse}')
printf'R-squared: {r2}')

plt.scatter(y_test, y_pred)


plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.title('Actual vs. Predicted Values')
plt.show()
```

Result:

```
Mean Squared Error: 20.40798465402493
R-squared: 0.9251346176444606
```

Evaluation Metrics are:

- ✓ Mean Squared Error
- ✓ R-squared



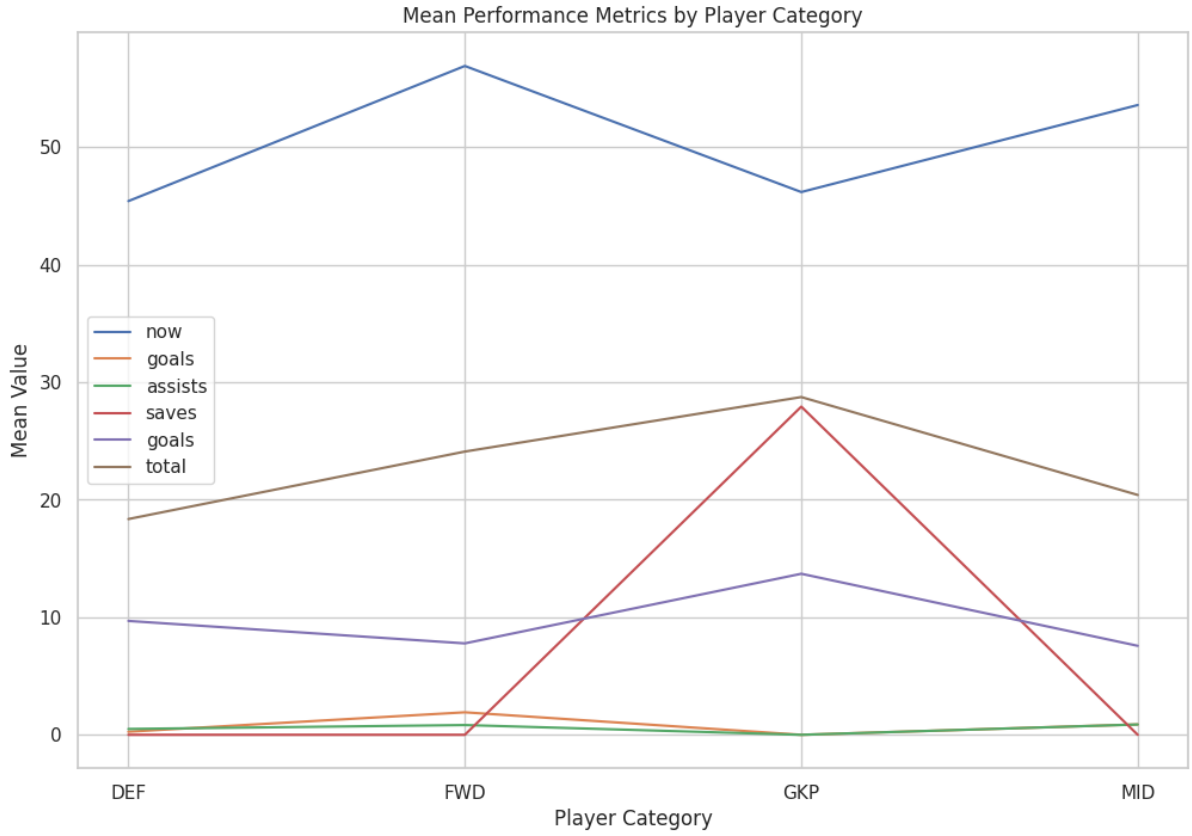
To evaluate the performance of the Machine Learning model – Regression Analysis a plot between the Predicted Values and Actual Values. We can see the linear relation between them. Majorly no deviation can be seen from one another.

5. Conclusions

For the undervalued players for the positions FWD, DEF, MID there are 19 players whose cost in the range of 40-50.

For the Goal Keepers there are 2 players who are undervalued with cost 39.

Below Multiple Line graph displays about the description of every attribute of the undervalued players (Our best players to buy).



now_cost

mean = 48.489906

min = 38.000000

25% = 44.000000

50% = 45.000000

75% = 50.000000

max = 140.000000

Results:

- ✓ Players with significant contributions exceeding the mean were identified.
- ✓ Undervalued players across different positions were highlighted.
- ✓ Goalkeepers with promising statistics were pinpointed for consideration.
- ✓ Visualizations provided intuitive insights into player performance.
- ✓ Machine learning analysis demonstrated potential for predictive modeling.

6. References

[1] Nicholas Bonello, Joeran Beel, Seamus Lawless, Jeremy Debattista. “Multi-stream Data Analytics for Enhanced Performance Prediction in Fantasy Football.” In 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. 2019.

[2] Pratik Pokharel, Arun Timalsina, Sanjeeb Panday, Bikram Acharya "Fantasy Premier League - Performance Prediction" Proceedings of 12th IOE Graduate Conference, 2022.

[3] Adrian Becker and Xu Andy Sun “An analytical approach for fantasy football draft and lineup management” J. Quant. Anal. Sports 2016.

[4] Benjamin Motz “Fantasy Football: A Touchdown for Undergraduate Statistics Education” 2013 Conference: Games+Learning+Society At: Madison, Wisconsin.

[5] Summers, A. E., T. B. Swartz, and R. A. Lockhart. 2007. Statistical Thinking in Sports. Chapman and Hall/CRC, chapter Optimal Drafting in Hockey Pools.