FC24 STRIKER ANALYSIS

Introduction: Overall Project Objective

Background: Strikers play a crucial role in scoring goals and shaping game outcomes. A striker's skill set includes Finishing Ability, Skill Moves, Sprint Speed, and Overall Rating. Understanding these attributes and how they influence performance can help clubs and scouts make informed decisions.
Dataset: The analysis uses the FC24 dataset, which includes player metrics such as Age, Foot Preference, Strength, and Agility. Analyzing patterns and correlations in this data can reveal insights into the qualities of top-performing strikers.

Objectives

Main Objective

- Data Exploration: Analyze distributions, patterns, and correlations in striker attributes.
- Contingency Analysis: Examine foot preference (left vs. right) across age groups to see if it impacts performance.
- Visualization: Use charts (e.g., histograms, box plots) to highlight key insights visually.
- Predictive Modeling: Identify which attributes are most strongly associated with high striker performance.

Secondary Objectives

- To provide valuable insights for clubs, analysts, and scouts to enhance team strategy and player recruitment by understanding the essential traits of a successful striker.

Data Visualisation and Descriptive Statistics

Player performance (as measured by overall rating) tends to peak between ages 26-30.

- **Boxplot of Overall Rating by Age Group**
  - **Median Ratings:** Players aged 26-30 have the highest median overall rating, suggesting players in this age group often perform at their peak.
  - **Spread of Ratings:** Age groups 15-20 and 36-40 show a narrower range of ratings, indicating more consistency. The 21-30 age groups, especially 26-30, show a broader range (IQR), suggesting more variability in player quality within these groups.
  - **Outliers:** All age groups have outliers, particularly high ratings, meaning there are standout players in each group.
  - **Trend:** Ratings generally increase with age until 26-30, then level off or slightly decrease in older age groups, indicating peak performance typically occurs in the late 20s.

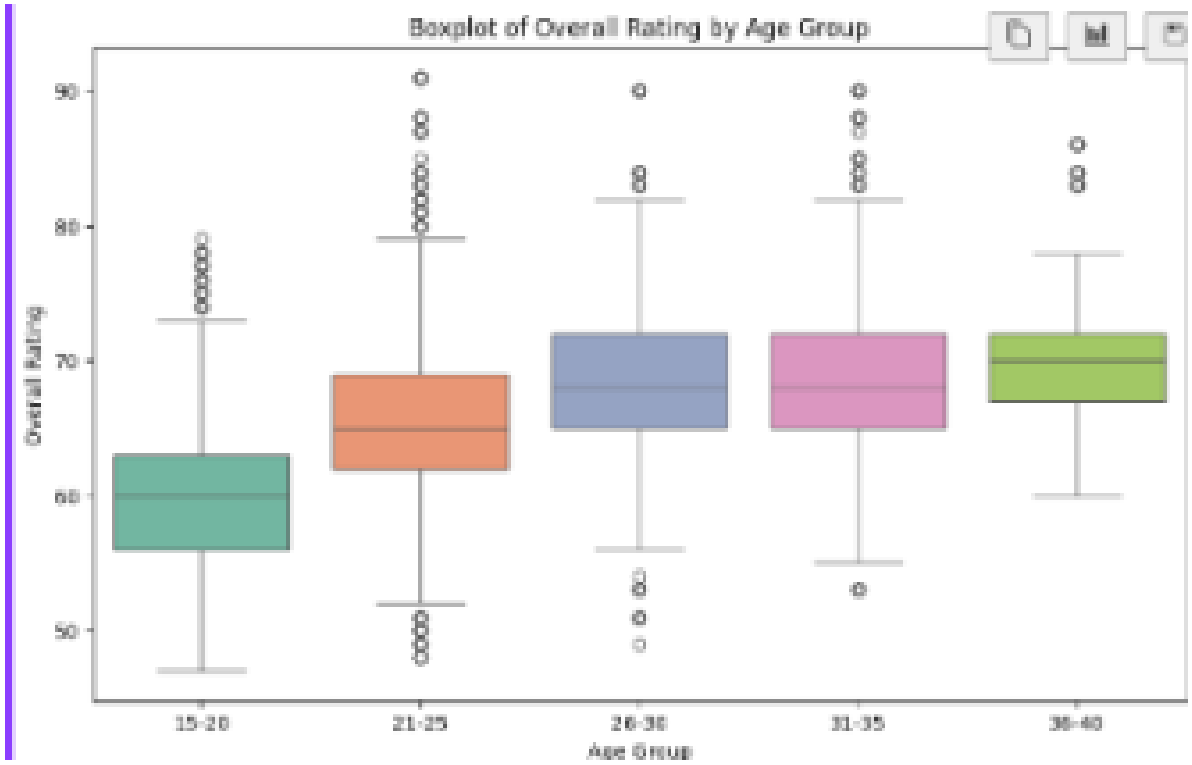


Figure: Boxplot of Overall Rating by Age Group

Average Overall Rating by Age Group

Players' overall ratings grow with age, peak at 26-30, and remain strong as players gain experience in their 30s.

- **Trend**:Average rating increases with age, peaking in the 26-30 group.
- **Peak Performance (26-30):** Players in this age group have the highest average rating, indicating that players often reach their peak performance in their late 20s.
- **Experienced Consistency (31-40):** Older groups (31-35 and 36-40) maintain high ratings, suggesting that experience helps sustain performance.
- **Young Players (15-20):** This group has the lowest average rating, reflecting skill development in the early stages of their careers.

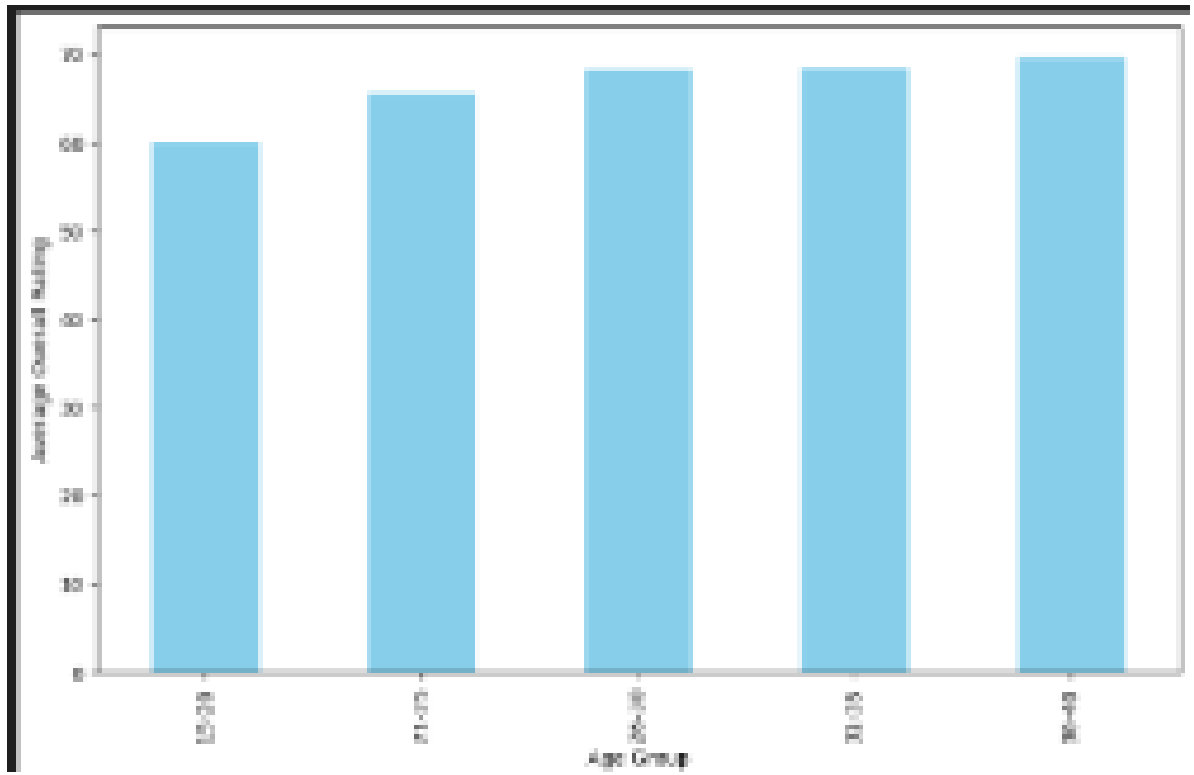
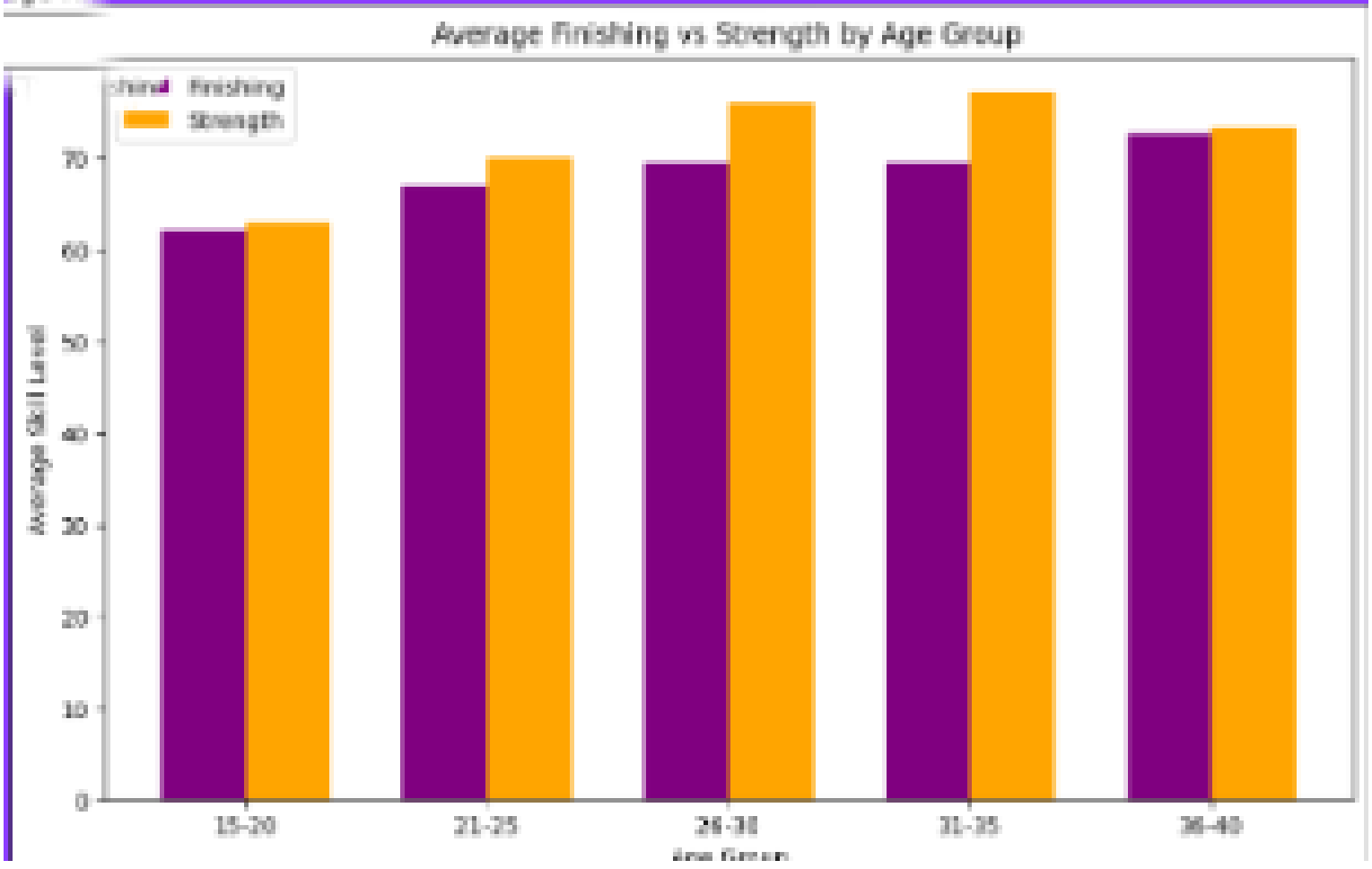


Figure: Average Overall Rating by Age Group

Average Finishing vs Strength by Age Group

Overview: Players typically reach optimal strength and finishing abilities in their late 20s and early 30s, with younger players still building these skills.

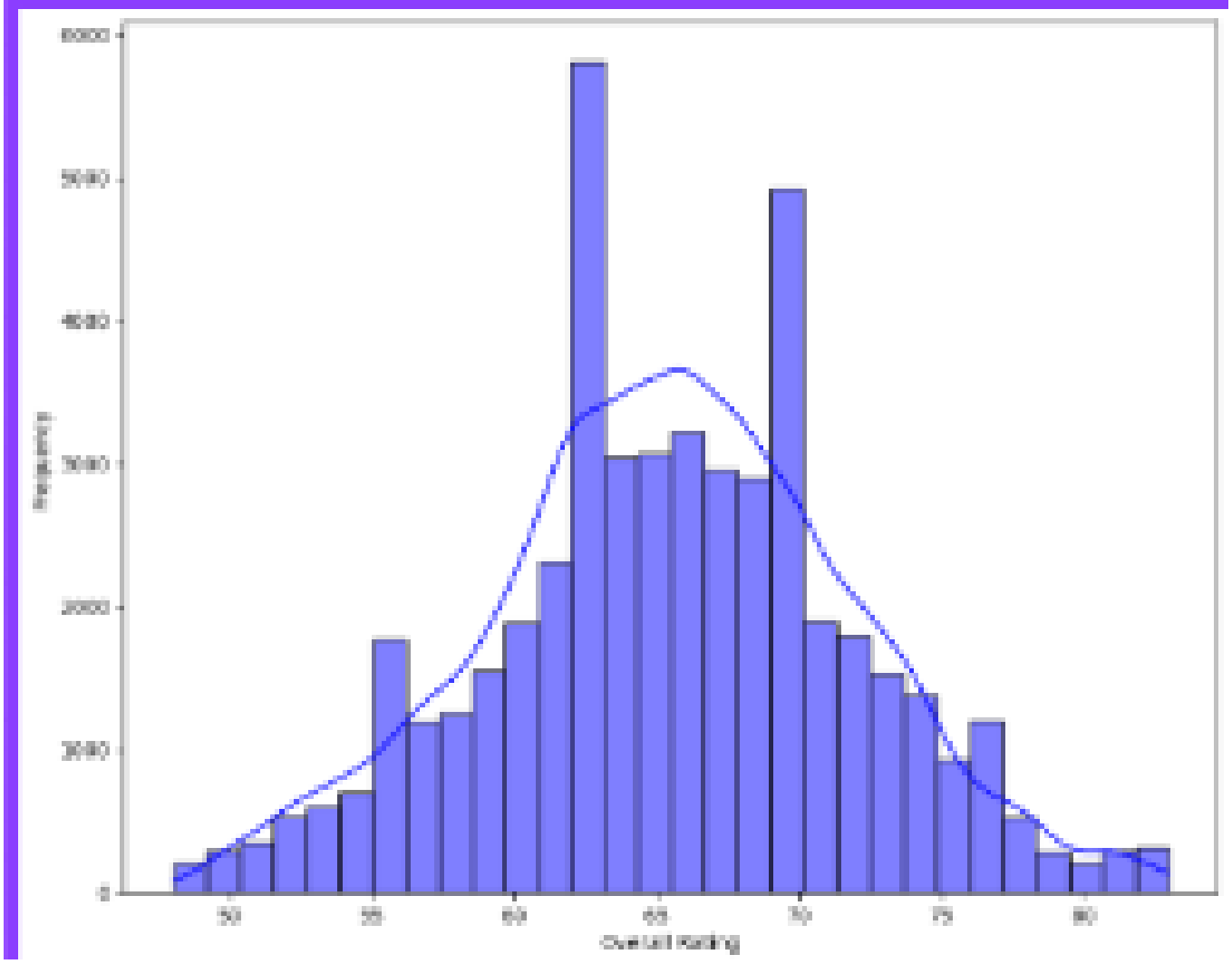
- **Strength:** Peaks in the 26-30 age group and remains high in the 31-35 group, indicating physical development with age.
- **Finishing:** Relatively consistent across age groups, with a slight increase in the 26-35 age range (prime scoring years).
- **Younger Players (15-20):** Display lower levels in both finishing and strength, reflecting ongoing skill and physical development.


Average Finishing vs Strength by Age Group
Finishing
Strength
Average Skill Level
0
10
20
30
40
50
60
70
15-20
21-25
26-30
31-35
36-40

Number of Left Footed vs Right Footed Players

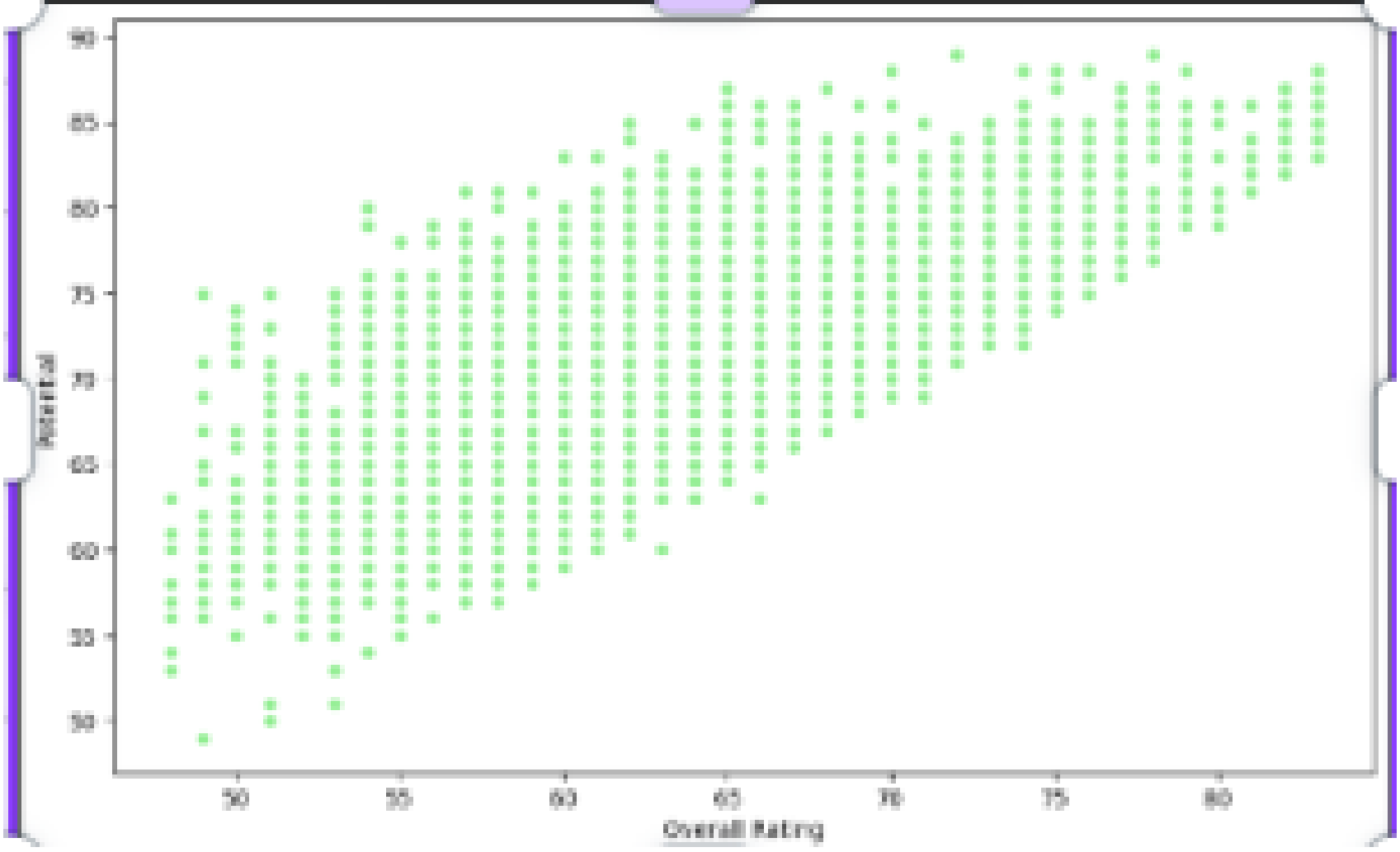
The data has a central tendency with mild skewness and specific rating concentrations, useful for player performance analysis.

- **Shape:** The histogram of Overall Ratings shows a roughly normal distribution, with a slight skew towards the right.
- **Peaks:** Noticeable peaks around 65 and 70, indicating common rating levels among players.
- **Range:** Ratings mostly fall between 50 and 80, with few players rated below 50 or above 80.
- **Insight):** The distribution suggests that most players have average ratings, with fewer highly or lowly rated players, aligning with expectations in a competitive dataset.


0
50
55
60
65
70
75
80
Overall Rating

Scatter Plot of Overall Rating vs Potential

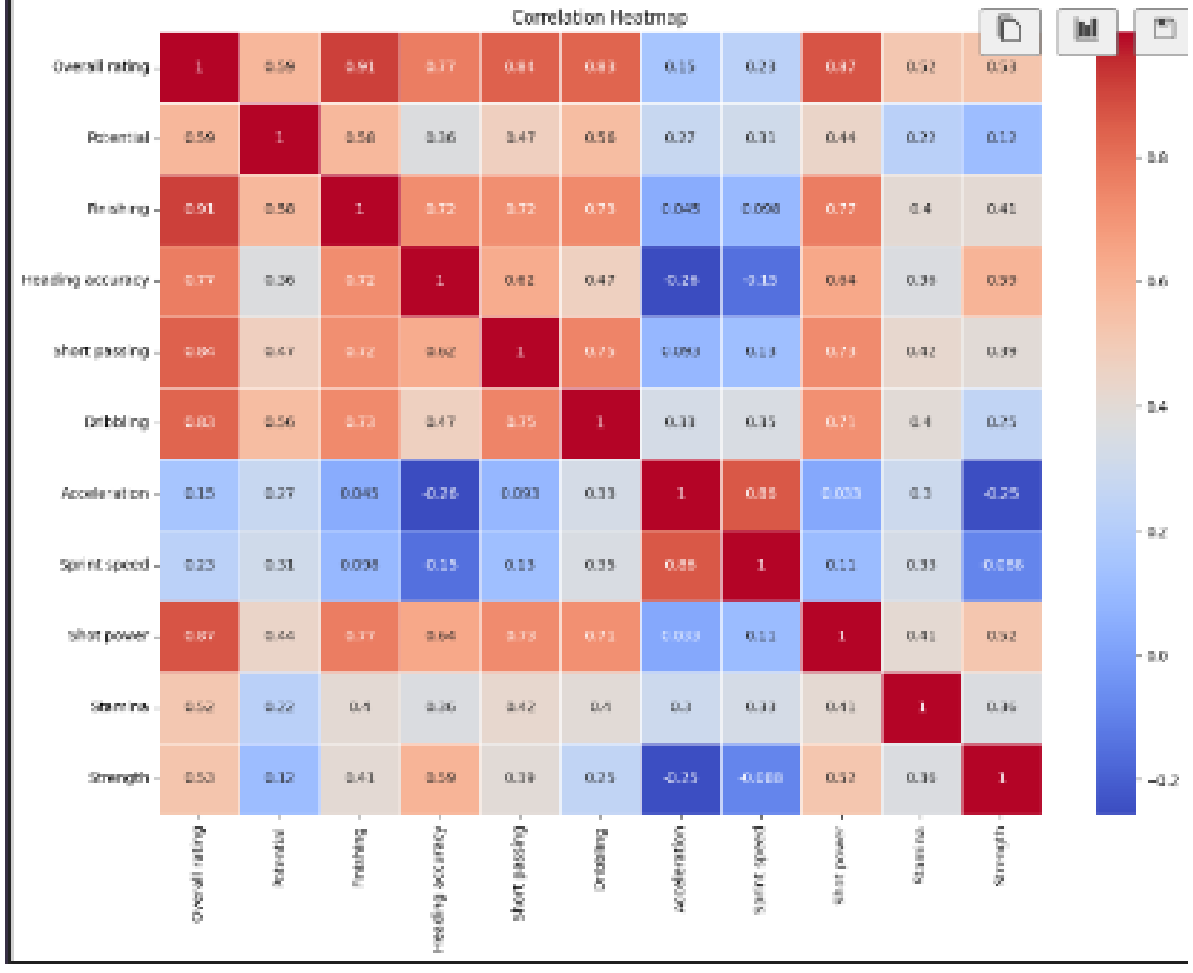
- **Positive Correlation:** As "Overall Rating" increases, "Potential" also tends to increase. This suggests that players with higher current ratings generally have higher potential.
- **Range of Ratings:** The "Overall Rating" spans from around 50 to 80, while "Potential" ranges from about 50 to 90.
- **Dense Clusters:** There are dense clusters of data points in the mid-range (Overall Rating between 60 and 70), which could indicate a common skill level in this dataset.


90
85
80
75
70
65
60
55
50
Potential
50
55
60
65
70
75
80
Overall Rating

Correlation Heatmap

The heatmap highlights which skills most contribute to overall performance, with strong connections in technical and speed-related attributes.

- **High Correlations:** "Overall Rating" has strong positive correlations with "Finishing" (0.91), "Shot Power" (0.87), and "Short Passing" (0.84), suggesting these attributes are key contributors to a high overall rating.
- **Notable Attribute Pairings:** "Sprint Speed" and "Acceleration" have a high correlation (0.86), indicating that players who are faster also tend to accelerate quickly. "Short Passing" and "Dribbling" are also strongly correlated (0.75).
- **Weak or Negative Correlations:** Some attributes, like "Acceleration" and "Heading Accuracy" (-0.26), show weak or negative correlations, suggesting these skills are less related.


Correlation Heatmap
Overall rating
Potential
Finishing
Heading accuracy
Short passing
Dribbling
Acceleration
Sprint speed
Stamina
Strength
1
0.59
0.91
0.77
0.81
0.83
0.15
0.23
0.87
0.52
0.53
0.59
1
0.58
0.36
0.47
0.58
0.27
0.31
0.44
0.23
0.12
0.91
0.58
1
0.72
0.72
0.73
0.045
0.068
0.77
0.4
0.41
0.77
0.36
0.72
1
0.62
0.47
-0.38
-0.15
0.64
0.38
0.33
0.81
0.47
0.72
0.62
1
0.73
0.093
0.13
0.73
0.42
0.39
0.83
0.58
0.73
0.47
0.75
1
0.33
0.35
0.71
0.4
0.25
0.15
0.27
0.045
-0.26
0.093
0.33
1
0.88
0.033
0.3
-0.25
0.23
0.31
0.038
0.15
0.13
0.35
0.88
1
0.13
0.33
0.038
0.87
0.44
0.77
0.64
0.73
0.71
0.033
0.13
1
0.41
0.62
0.52
0.22
0.4
0.26
0.42
0.4
0.3
0.33
0.41
1
0.46
0.53
0.12
0.41
0.59
0.39
0.25
-0.25
-0.038
0.62
0.38
1
0.8
0.6
0.4
0.2
0.0
−0.2

Contingency Table for Age Group and Dominant Foot

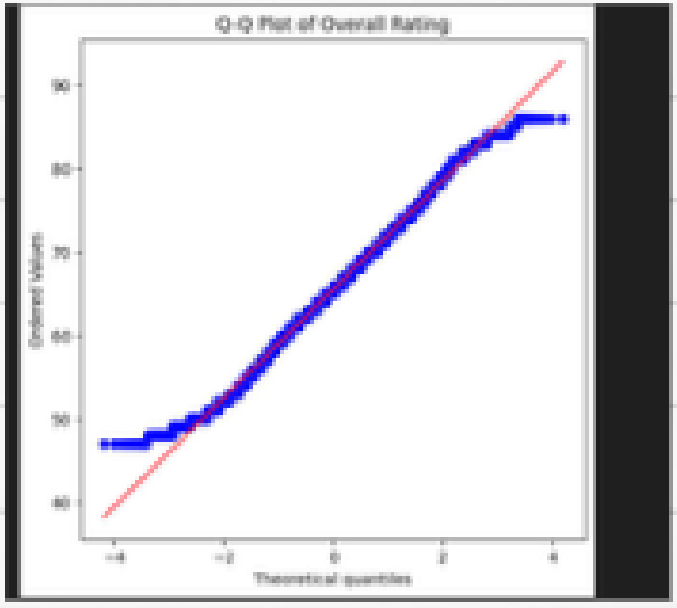
This distribution shows a consistent trend in foot preference with a clear peak in the 21-25 age group.

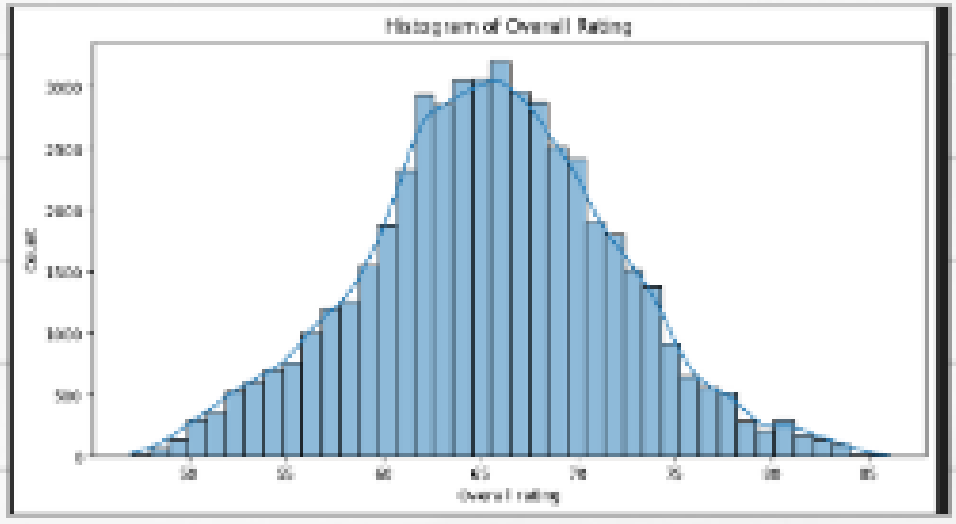
- Right-footed players are more common across all age groups.
- Age Group 21-25 has the highest count of players, for both left and right-footed categories.

Probability Distribution

1 QUANTILES OF OVERALL RATING

- **Normality Check:** The data mostly follows the red line, indicating that "Overall Rating" is approximately normally distributed.
- **Deviation at Extremes:** There are deviations at both lower and higher ends, suggesting some skewness or outliers in the distribution. Lower ratings (left tail) and higher ratings (right tail) show slight departures from normality.
- **Middle Range Fit:** The middle quantiles closely align with the theoretical quantiles, confirming that the majority of "Overall Rating" values fit a normal distribution pattern.
- **The Jarque-Bera statistic of approximately 5.8023 indicates that the skewness and kurtosis of the Overall rating distribution differ significantly from a normal distribution.**
- **The p-value of 0.05496 is close to 0.05 so its fail to reject the null hypothesis that the distribution is normal.**






```
jb_statistic, p_value = stats.jarque_bera(st_players['Overall rating'])

print(f"JB Statistic: {jb_statistic}")
print(f"P-value: {p_value}")
```

✓ 0.0s

```
JB Statistic: 5.802259948439867
P-value: 0.05496108035090205
```

2.1 Density of Overall Rating by Foot Preference

- **Statistical Results:**
  - **T-Statistic:** 4.61
  - **P-Value:** 4.13×10^{-6}
- **Interpretation:**
  - **Significant Effect:** The p-value is far below typical significance levels (e.g., 0.05).
  - **Reject the Null Hypothesis:** Evidence suggests the effect is unlikely due to chance.
  - **Conclusion:** The result is statistically significant, indicating a meaningful effect or relationship.

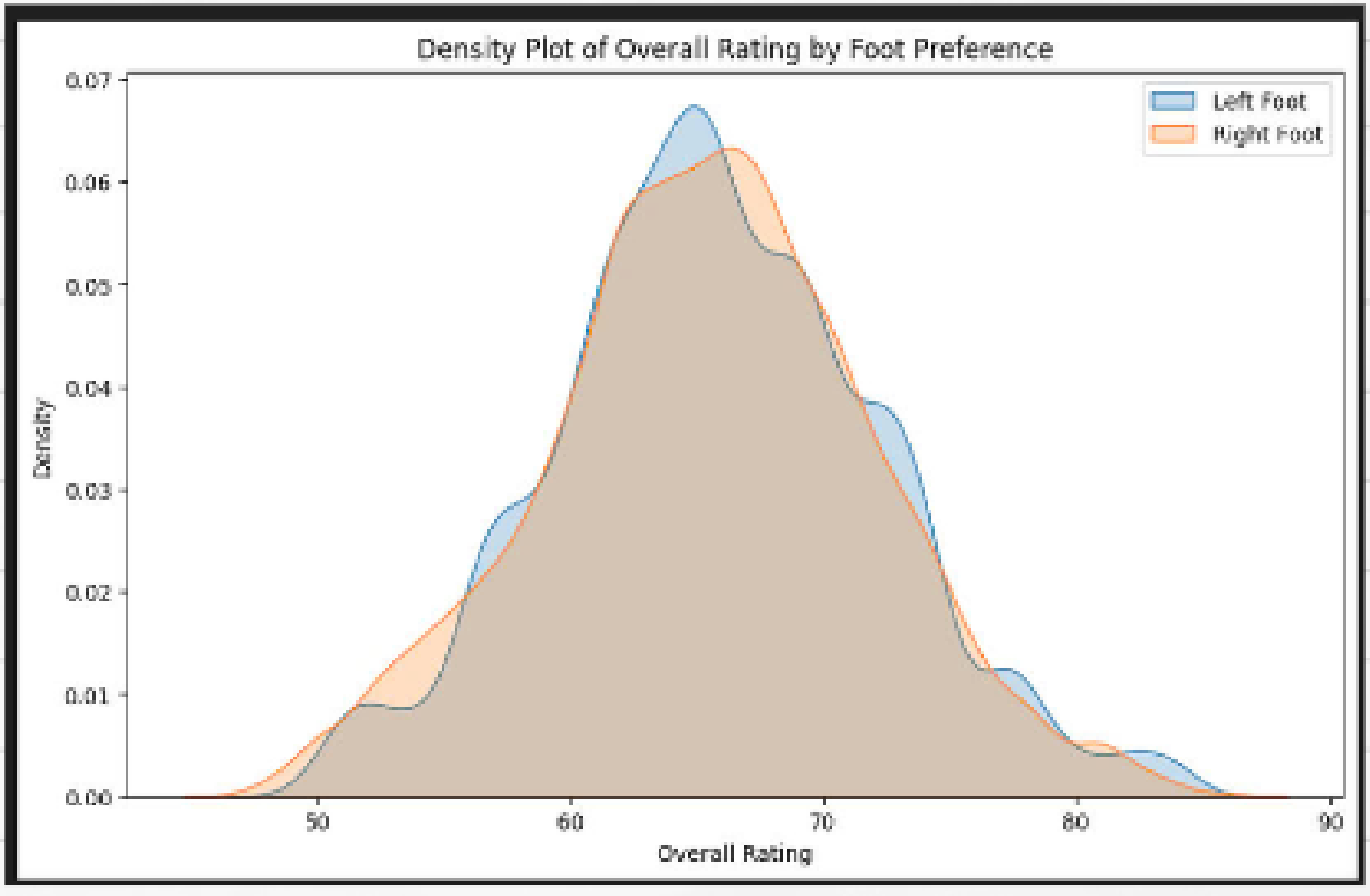
```
left_footers = st_players[st_players['foot_Left'] == 1]['Overall rating']
right_footers = st_players[st_players['foot_Right'] == 1]['Overall rating']

t_statistic, p_value = stats.ttest_ind(left_footers, right_footers, equal_var=False)

print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")
```

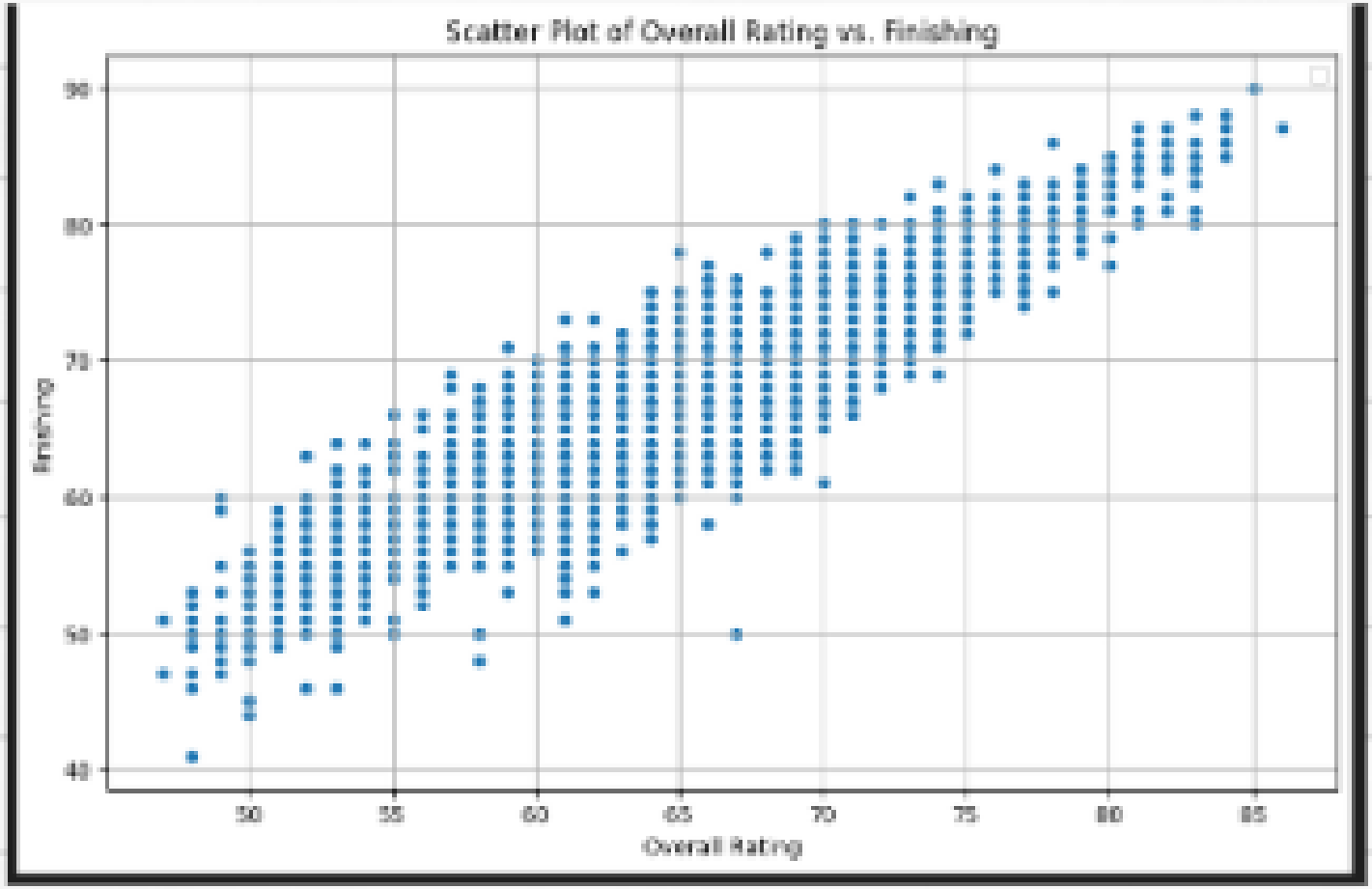
✓ 0.0s

```
T-statistic: 4.007138825055766
P-value: 6.127725790868114946-05
```



2.2 Density of Overall Rating by Foot Preference

- **The Pearson correlation coefficient of 0.9163 indicates a very strong positive linear relationship between the two variables being analyzed.**
- **Correlation: The scatter plot show there is a correlation between Overall rating and Finishing. The points tend to cluster along a straight line, it indicates a strong correlation. A positive slope suggests that higher overall ratings correspond to higher finishing scores.**




```
correlation_coefficient, p_value = stats.pearsonr(st_players['Finishing'], st_players['Overall rating'])

print("Hệ số tương quan Pearson:", correlation_coefficient)
print("p-value:", p_value)
```

✓ 0.0s

```
Hệ số tương quan Pearson: 0.9163559110868734
```

multipple linear regression

ADVANTAGES

- **01:** Captures Complexity: Striker performance is influenced by multiple attributes simultaneously. Multiple linear regression enables you to include all these factors to see how they jointly impact the target variable.
- **02:** Identifies Interactions: With multiple predictors, you can test whether specific combinations of skills might influence the performance, something you wouldn't capture with a single predictor.
- **03:** Quantifies Individual Impact: It not only shows if each attribute is influential but also quantifies how much each one contributes, holding other factors constant.
- **04:**Prediction Power: Multiple linear regression helps make more accurate predictions about striker performance based on their combined attributes, which can be useful in player ranking or scouting.

- **Mean Squared Error (MSE)**: The MSE of 0.45096 is quite low. MSE represents the average squared difference between the actual and predicted values, so a lower value indicates that your model's predictions are generally close to the actual data points. In this case, the small MSE shows that your model is making very accurate predictions.
- **R^2 Score**: An R^2 score of 0.9893 indicates that about 98.93% of the variability in striker performance (your target variable) is explained by the model. This is a very high R^2 value, suggesting that the independent variables chosen are highly effective at predicting the target variable and that the model has an excellent fit.
- **Overall Interpretation**
  - **Model Accuracy:** The low Mean Squared Error (MSE) indicates our model predictions are close to actual performance data, demonstrating high accuracy.
  - **Model Fit:** With an R^2 score of 98.93%, nearly all variability in striker performance is explained by the model, showing an excellent fit.
  - **Key Takeaway:** The multiple linear regression model effectively predicts striker performance based on selected

attributes, though further testing is recommended to confirm its reliability on unseen data.

Conclusion

Strengths:

- **Large Sample Size: With a dataset of 353,198 players, the study benefits from a large sample size, which enhances the reliability of the statistical analysis. We can still extend the problem to other positions in this dataset.**
- **Strong Correlation: Most of the important indicators have high correlation coefficients, which ensures that the model gives good results.**
- **High accuracy: The combination of a low MSE and a high R^2 score indicates that the model is performing exceptionally well. It suggests that the model's predictions are accurate and that it captures the underlying patterns in the data effectively.**

- **Limitation:**
  - **Changes in real time: Because this is data about players, the data will change continuously, so it may affect future parameters, thereby reducing model performance.**
  - **Many outliers: Because there are not too many famous players, most of the players with Overall rating ¿ 86 will affect and make this no longer a normal distribution, so we have to delete them.**
- **Future Directions:**
  - **Continue with other statistic and position**
  - **Find more data to use the model**
  - **Apply the model to future changes to see the difference**