Soccer Player's Salary Prediction Insights

Problem Description

Using data found online attempting to predict soccer player's salaries based on player statistics. It is interesting to look at players' pay from different angles, like position, club, location, etc. And if I can get a model that works, it would be very exciting to predict player salaries based on their statistics.

Research Questions I came up with

1. What player statistics help to predict a player's salary?
2. What factors influence a player's goal-scoring ability?
3. Is there variance across different leagues in Europe and USA? How do different playing positions across clubs affect salary?

Data

This is a dataset with information on soccer players' performance from top clubs around the world. I will be using categorical and continuous inputs to predict the salary of each player. There are 225 rows and 47 columns. For a full list and description of the variables, please see **Appendix A - Variable Descriptions.**

To explore the data I used Python and Tableau. I looked at outliers, top and bottom values, distributions, histograms, correlations, and descriptive statistics of the continuous and categorical variables.

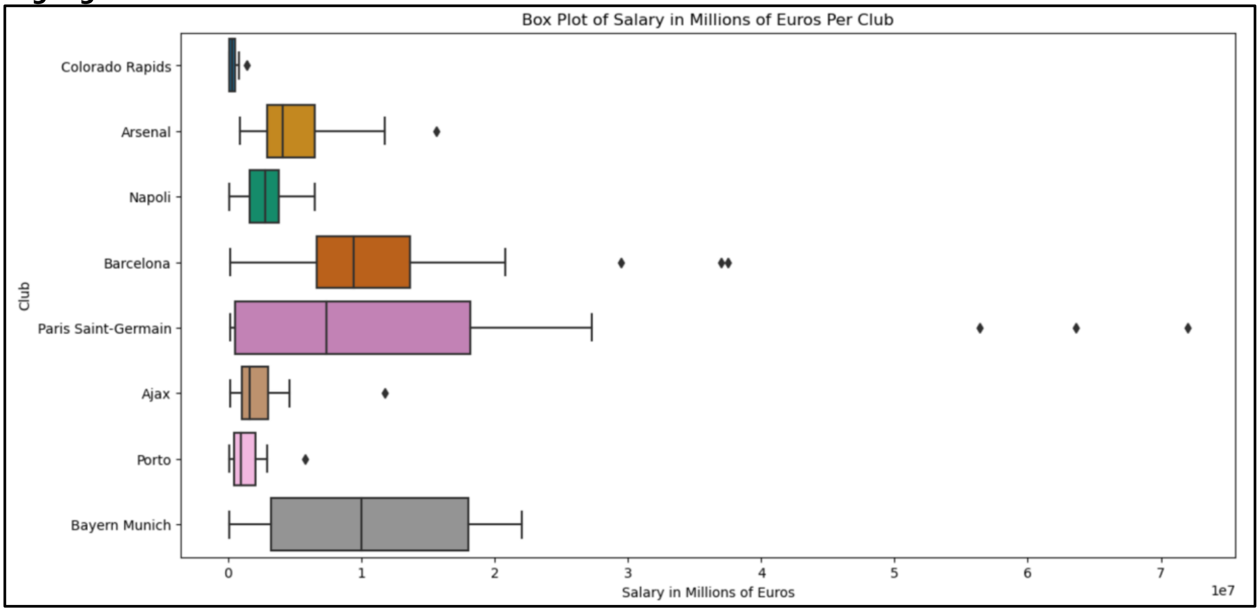
Python EDA

Here is a pdf link to the EDA work done in Python. Below are the overall notes and three highlights from the exploration:

EDA Notes:

- There are 204 rows, 47 columns
- Some nulls in the column Position_2. I can delete this column and concern theselves with Position_1. Alternatively, I can make it a y/n column and put a 'y' for everyone with a second position and a 'no' for people without.
- Output variable is not normally distributed.
- Lots of highly correlated data, will need to deal with this when creating models, perhaps conducting PCA analysis
- Only 4 positions are used in the dataset, which makes it ideal for modeling and analyzing the data.
- Since input 0s for all the missing data, it makes all the histograms not normally distributed. This is because, for example, goalies don't typically score goals, so their shots on target are 0, but that does not mean the stat is not right for analysis. This is the same for all positions, the stats that are important for their position are full, but the rest might be null or 0. This could pose challenges for modeling.
- Paris has some highly-paid players and superstars. I might want to remove those from the modeling set.

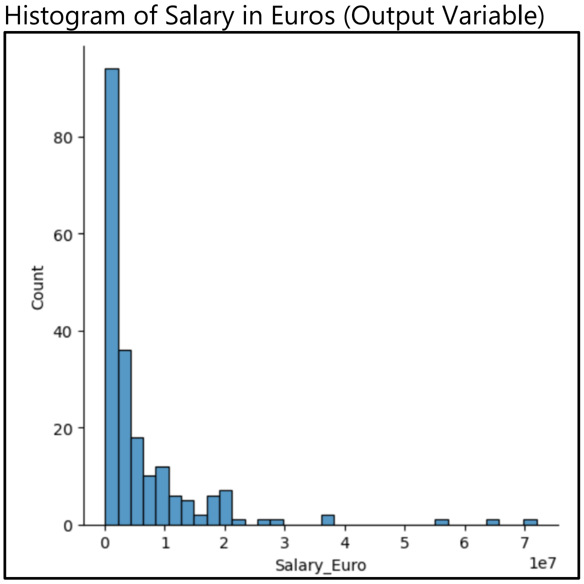
Highlight 1:



This is a very interesting and informative box plot. It really shows the salary outliers for Paris Saint Germain. Those players are Neymar, Messi, and Mbappe. All super famous futbol stars. It shows the difference betIen European teams and U.S. teams. The Colorado Rapids clearly make so much less than the EU teams.

Highlight 2:

Histogram of Salary in Euros (Output Variable)



This chart is important because it shows us that the output variable is not normally distributed. This is something I have to pay attention to when creating the models.

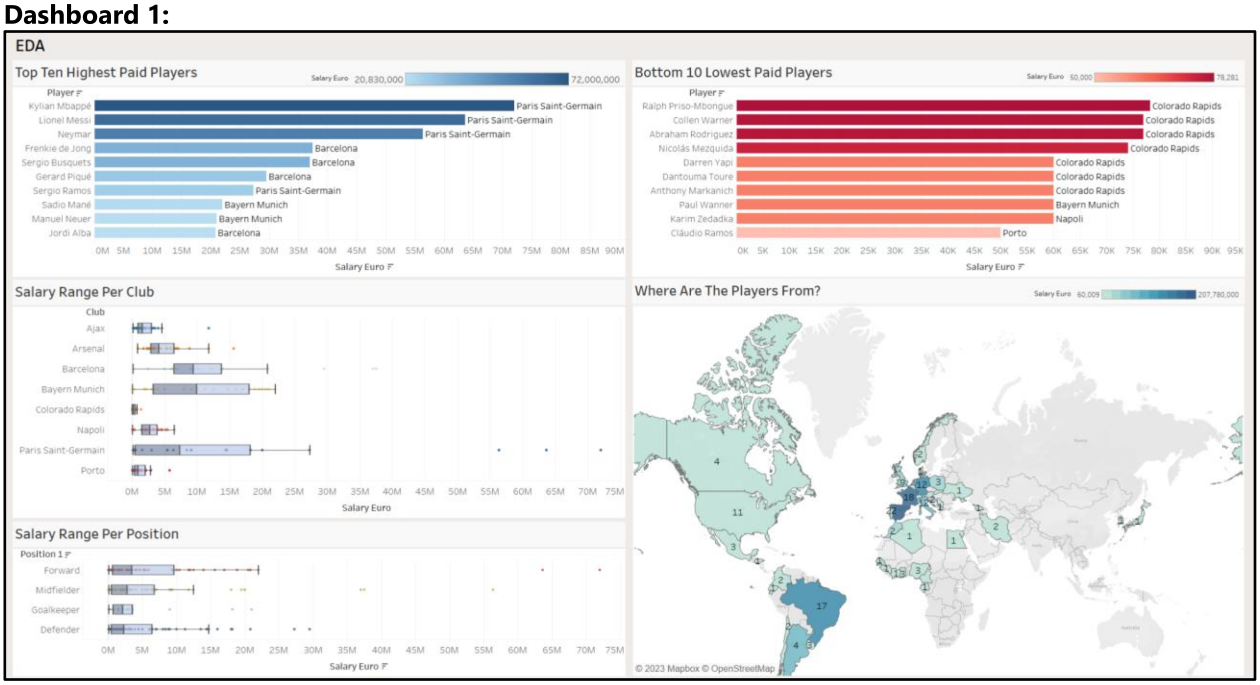
Highlight 3:

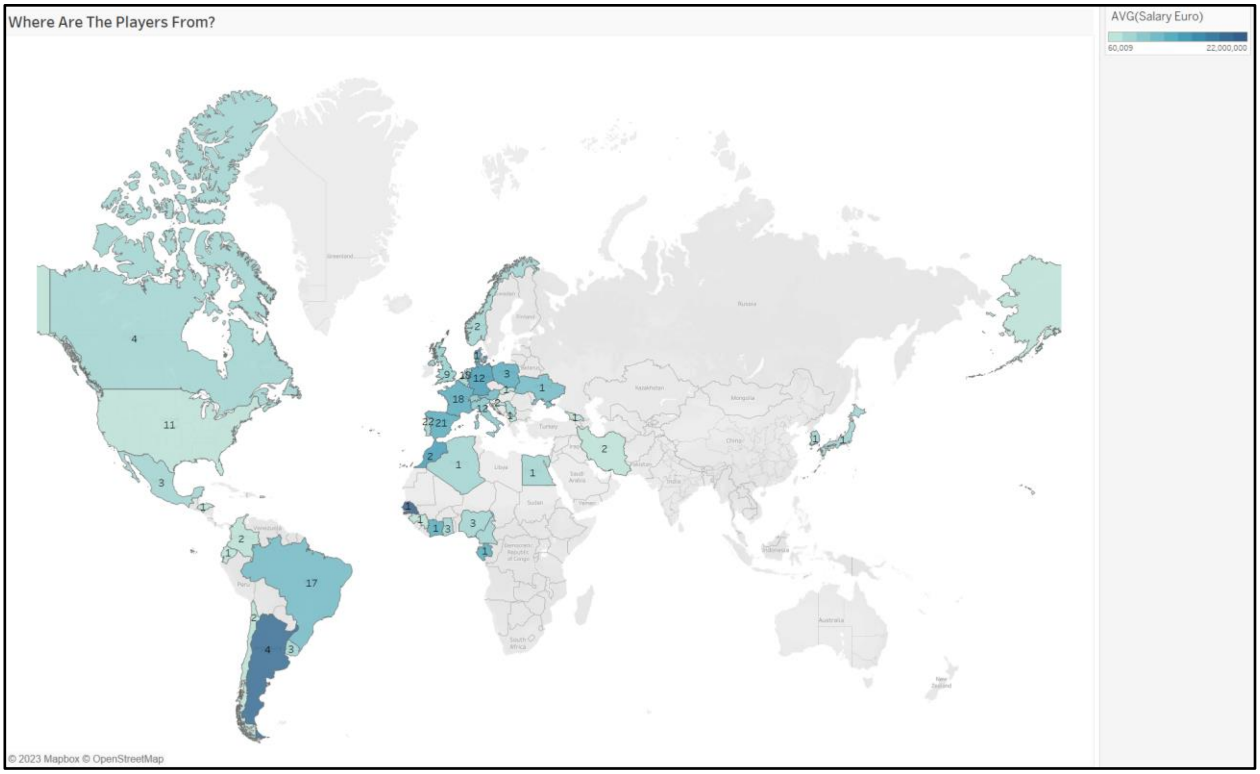
This chart makes a case for using modeling techniques to account for groups level predictions. You can see here that forwards tend to make the most money and goalies the least.

Tableau EDA

In Tableau I Ire able to quickly visualize different aspects of the dataset. I created maps, charts, and scatter plots to showcase the highlights from the EDA.

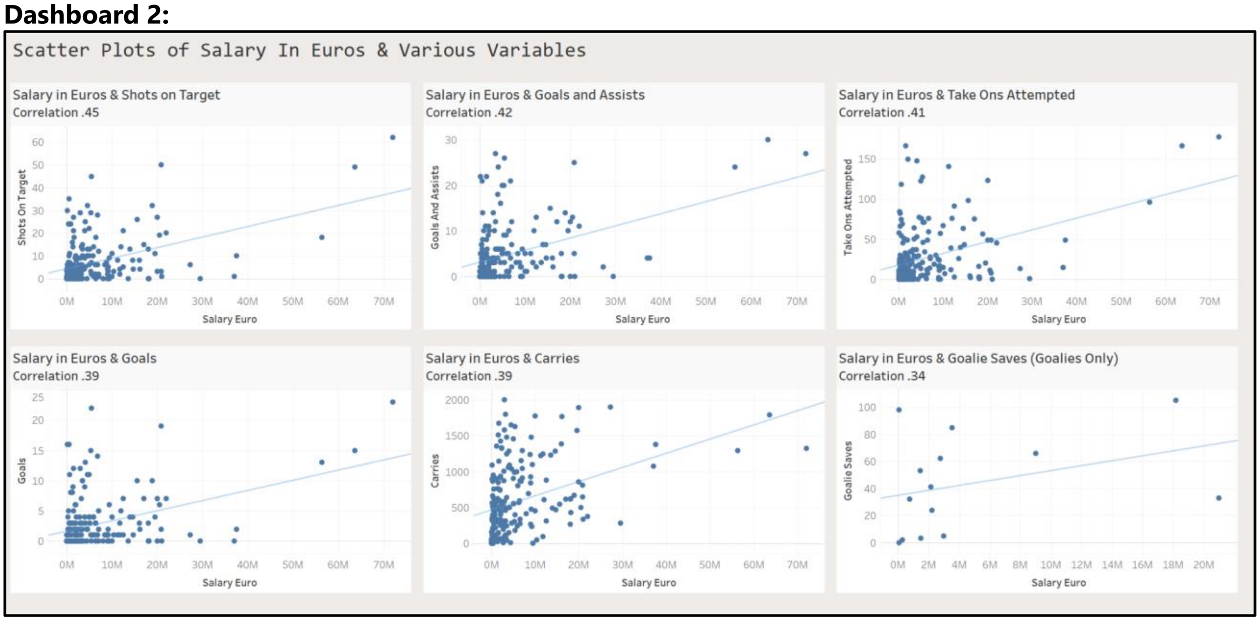
Dashboard 1:





The map helps to visualize where players are coming from. The darker the nation, the higher the average salary.

Dashboard 2:



These are the highest correlated variables with the output variable, *Salary* in Euros. The outliers are visible here as Ill as how most of the data is skeld to the left. The majority of salaries are betlen 0-20 million Euros.

Analysis of Questions

What player statistics help to predict a player's salary?

```
...                          OLS Regression Results
    ==============================================================================
    Dep. Variable:            Salary_Euro   R-squared:                       0.644
    Model:                            OLS   Adj. R-squared:                  0.614
    Method:                 Least Squares   F-statistic:                     21.19
    Date:                Tue, 30 May 2023   Prob (F-statistic):           8.10e-34
    Time:                        22:53:12   Log-Likelihood:                -3468.0
    No. Observations:                 204   AIC:                             6970.
    Df Residuals:                     187   BIC:                             7026.
    Df Model:                          16
    Covariance Type:            nonrobust
    ===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
    ---------------------------------------------------------------------------------------------------
    Age                          4.775e+05   9.61e+04      4.969      0.000    2.88e+05    6.67e+05
    Goals                       -6.362e+05   3.01e+05     -2.111      0.036   -1.23e+06   -4.17e+04
    Assists                       3.71e+05   2.84e+05      1.307      0.193   -1.89e+05    9.31e+05
    Red_Cards                     4.39e+06   1.33e+06      3.301      0.001    1.77e+06    7.01e+06
    Progressive_Passes           2.347e+04   8134.494      2.886      0.004    7427.435    3.95e+04
    Progressive_Passes_Received -2.861e+04   1.02e+04     -2.801      0.006   -4.88e+04   -8460.205
    Shots_Attempted             -4.514e+05   8.34e+04     -5.410      0.000   -6.16e+05   -2.87e+05
    Shots_On_Target              1.541e+06    2.2e+05      7.005      0.000    1.11e+06    1.97e+06
    Take_Ons_Attempted            9.03e+04   2.33e+04      3.875      0.000    4.43e+04    1.36e+05
    Club_Arsenal                -4.454e+06   1.34e+06     -3.321      0.001    -7.1e+06   -1.81e+06
    Club_Barcelona              -1.946e+06   1.43e+06     -1.357      0.176   -4.77e+06    8.83e+05
    ...
    Notes:
    [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
    [2] The smallest eigenvalue is 1.85e-32. This might indicate that there are
    strong multicollinearity problems or that the design matrix is singular.
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

The OLS regression results indicate that the overall model explains **64.4%** of the variation in the dependent variable, Salary_Euro, as indicated by the R-squared value. The adjusted R-squared value, which accounts for the number of predictors in the model, is **61.4%**.

The following predictors Ire significant for predicting Salary, having p-values of less than 0.05:

- Age
- Red_Cards
- Progressive_Passes
- Progressive_Passes_Received
- Shots_Attempted
- Shots_On_Target
- Take_Ons_Attempted
- Club_Arsenal
- Club_Colorado Rapids
- Club_Napoli
- Club_Porto Governing_Country_Germany
- Governing_Country_Italy
- Governing_Country_Netherlands
- Governing_Country_Portugal

Variables such as Goals, Assists, Club_Barcelona, Club_Bayern Munich, and Club_Paris Saint-Germain are not statistically significant at the 0.05 level.

The regression model also indicates potential multicollinearity issues or a singular design matrix, as suggested by the condition number and the smallest eigenvalue.

Based on these results, it is recommended to consider the statistically significant predictors in determining player salaries. Variables such as age, red cards, progressive passes, shots attempted on target, take-ons attempted, and club and governing country affiliations can be used as factors in salary negotiations.

Here is the interpretation of each variable's coefficient in relation to the target variable, **Salary_Euro**:

- **Age:** For every unit increase in Age, there is an increase of approximately 477,500 Euros in Salary_Euro, holding other variables constant. This suggests that older players have higher salaries.
- **Goals:** For every unit increase in Goals, there is a decrease of approximately 636,200 Euros in Salary_Euro, holding other variables constant. This indicates that scoring more goals may not necessarily lead to higher salaries.
- **Assists:** For every unit increase in Assists, there is an increase of approximately 371,000 Euros in Salary_Euro, holding other variables constant. This suggests that players who provide more assists tend to have higher salaries.
- **Red_Cards:** For every unit increase in Red_Cards, there is an increase of approximately 4,390,000 Euros in Salary_Euro, holding other variables constant. This unexpected result implies that players who receive more red cards have higher salaries, which could be due to other factors related to their playing style or reputation.
- **Progressive_Passes:** For every unit increase in Progressive_Passes, there is an increase of approximately 23,500 Euros in Salary_Euro, holding other variables constant. This indicates that players who make more progressive passes tend to have higher salaries.
- **Progressive_Passes_Received:** For every unit increase in Progressive_Passes_Received, there is a decrease of approximately 28,600 Euros in Salary_Euro, holding other variables constant. This suggests that players who receive more progressive passes may have loIr salaries because they are less involved in playmaking.
- **Shots_Attempted:** For every unit increase in Shots_Attempted, there is a decrease of approximately 451,400 Euros in Salary_Euro, holding other variables constant. This indicates that players who attempt more shots may not necessarily command higher salaries.
- **Shots_On_Target:** For every unit increase in Shots_On_Target, there is an increase of approximately 1,541,000 Euros in Salary_Euro, holding other variables constant. This suggests that players with more shots on target tend to have higher salaries.
- **Take_Ons_Attempted:** For every unit increase in Take_Ons_Attempted, there is an increase of approximately 90,300 Euros in Salary_Euro, holding other variables constant. This implies that players who attempt more take-ons tend to have higher salaries.
- **Club variables (e.g., Club_Arsenal, Club_Barcelona):** The coefficients represent the salary difference compared to a reference club (likely omitted from the model). Positive coefficients indicate higher salaries compared to the reference club, while negative coefficients suggest loIr wages.
- **Governing_Country variables (e.g., Governing_Country_Germany):** The coefficients represent the salary difference compared to a reference governing country (likely omitted from the model).

Positive coefficients indicate higher salaries compared to the reference country, while negative coefficients suggest loIr salaries.

What factors influence a player's goals scoring ability?

From an offensive perspective, clubs and scouts often first look at high-scoring players. With this in mind, I constructed a Poisson model to determine what factors influence goal scoring, so that scouts and clubs know what to look for.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  Goals   No. Observations:                  204
Model:                            GLM   Df Residuals:                      196
Model Family:                 Poisson   Df Model:                            7
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -334.13
Date:                Tue, 30 May 2023   Deviance:                       314.07
Time:                        22:53:31   Pearson chi2:                     278.
No. Iterations:                     5
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   0.3692      0.298      1.238      0.216      -0.215       0.954
Age                    -0.0284      0.012     -2.379      0.017      -0.052      -0.005
Assists                 0.0776      0.015      5.125      0.000       0.048       0.107
Yellow_Cards            0.0378      0.021      1.828      0.068      -0.003       0.078
Red_Cards              -0.0920      0.138     -0.668      0.504      -0.362       0.178
Expected_Goals          0.1605      0.009     18.146      0.000       0.143       0.178
Progressive_Carries    -0.0006      0.002     -0.354      0.723      -0.004       0.003
Touches                 0.0002   8.92e-05      2.712      0.007    6.71e-05       0.000
=========================================================================================
```

<u>Variables used in the model:</u>

- **Age:** is known to have an impact on a player's performance and goal-scoring ability. Younger players might have more energy and agility, which could contribute to higher goal-scoring rates.
- **Assists:** Assists can be a good indicator of a player's involvement in the attacking play and their ability to create goal-scoring opportunities for themselves and their teammates.
- **Yellow_Cards:** While yellow cards might seem unrelated to goal-scoring, they could be indicative of a player's aggression and competitiveness, which might influence their goal-scoring ability.
- **Red_Cards:** Similar to yellow cards, red cards could indicate a player's aggressiveness and could potentially impact their goal-scoring opportunities if they receive suspensions or bans.
- **Expected_Goals:** Expected goals are a metric that quantifies the quality of scoring opportunities a player has had. Including this variable helps capture the player's goal-scoring potential based on the quality of chances they have created or received from being in ideal positions.
- **Progressive_Carries:** Progressive carries measure the ability of a player to carry the ball forward and advance the attack. Players with higher progressive carries have a better chance of getting into goal-scoring situations.
- **Touches:** The number of touches a player has on the ball could reflect their involvement in the game and their ability to handle the ball.

Overall, younger players, those with more assists, higher expected goals, and more touches tend to score more goals. Clubs should focus on recruiting or retaining younger players with high goal-scoring potential, such as having a high number of assists. Additionally, enctheaging players to be more involved in the game and increase their number of touches could lead to better goal-scoring opportunities.

It is important to note that other variables not included in this model could also impact goal scoring, such as playing position, playing time, and the quality of the opposition. Further analysis incorporating these variables could provide a more comprehensive understanding of goal-scoring in football.

<u>Odd ratios for the variables in the Poisson GLM model:</u>

- **Age:** The odds of scoring a goal decrease by a factor of exp(-0.0284) = 0.9719 for a one-year increase in age.
- **Assists:** The odds of scoring a goal increase by a factor of exp(0.0776) = 1.0808 for a one-unit increase in assists.
- **Yellow_Cards:** The odds of scoring a goal increase by a factor of exp(0.0378) = 1.0384 for a one-unit increase in yellow cards.
- **Red_Cards:** The odds of scoring a goal decrease by a factor of exp(-0.0920) = 0.9127 for a one-unit increase in red cards.
- **Expected_Goals:** The odds of scoring a goal increase by a factor of exp(0.1605) = 1.1746 for a one-unit increase in expected goals.
- **Progressive_Carries:** The odds of scoring a goal decrease by a factor of exp(-0.0006) = 0.9994 for a one-unit increase in progressive carries.
- **Touches:** The odds of scoring a goal increase by a factor of exp(0.0002) = 1.0002 for a one-unit increase in touches.

Is there variance across different leagues in Europe and USA? How do different positions across clubs affect salary?

Model:	MixedLM	Dependent Variable:	Salary_Euro
No. Observations:	204	Method:	REML
No. Groups:	8	Scale:	71884860372409.5938
Min. group size:	24	Log-Likelihood:	-3537.5747
Max. group size:	28	Converged:	Yes
Mean group size:	25.5		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	6166722.143	1973510.104	3.125	0.002	2298713.417	10034730.870
Club ID Var	28330787843437.816	2001583.956				

Group-level Variance Calculation:

```
28330787843437.777/(28330787843437.777+71884860372409.5938)
```

The estimated group-level variance indicates that approximately 28% of the variation in Salary_Euro is due to differences betlen clubs. The remaining 72% is attributed to individual-level factors. The intercept coefficient is significant, indicating a significant average difference in Salary_Euro across all clubs. Considering the hierarchical structure of the data, a Linear Mixed Effects Model is appropriate to account for both group-level and individual-level effects. This provides a comprehensive understanding of the factors influencing *Salary*.

Variance:

"Club ID Var": The estimated variance associated with the different clubs in relation to Salary_Euro is approximately 4.05 x 10^28 (405,406,207,681,180,625.0). This indicates a significant variability betlen clubs regarding their impact on *Salary*.

Covariance:

"Club ID x C(Position_1)[T.Forward] Cov": The covariance betlen the random effects of different clubs and the forward position is approximately 7.37. This suggests a moderate association betlen certain clubs and Salary_Euro for forward players.

"C(Position_1)[T.Forward] x C(Position_1)[T.Goalkeeper] Cov": The covariance betlen the random effects of forward and goalkeeper positions is approximately 6.76 x 10^12 (6,759,491,057,657.06). This indicates a positive association or overlap in Salary_Euro betlen these positions.

"Club ID x C(Position_1)[T.Midfielder] Cov": The covariance betlen the random effects of different clubs and the midfielder position is approximately 11.02. This suggests a certain level of association betlen certain clubs and Salary_Euro for midfielders.

"C(Position_1)[T.Goalkeeper] x C(Position_1)[T.Midfielder] Cov": The covariance betlen the random effects of goalkeeper and midfielder positions is approximately 9.57 x 10^11 (956,695,740,134.065). This indicates a potential association or interaction betlen these positions in relation to Salary_Euro.

These variance components and covariances provide insights into the variability and relationships betlen different groups (clubs) and positions (forward, goalkeeper, and midfielder) in the model. They help quantify the extent of variability and potential associations, contributing to a more comprehensive understanding of the mixed-effects model.

Ajax versus FC Barcelona Example:

- Ajax Defenders: The mean salary for Ajax's defenders is approximately 1.33 million.
- Barcelona Defenders: The mean salary for Barcelona's defenders is approximately 9.98 million.
- Interpretation: Barcelona's defenders have a significantly higher mean salary compared to Ajax's defenders, indicating that Barcelona may have invested more in their defenders or have higher-valued players in that position.

Conclusion

The first OLS regression model suggests that predictors including Age, Red_Cards, Progressive_Passes, Shots_Attempted, Shots_On_Target, Take_Ons_Attempted, and various club and governing country affiliations are the most significant predictors in determining player salaries.

Additionally, I saw that younger players, those with more assists, and more technical skills (e.g. touches) tend to score more goals or contribute to goal scoring. Clubs should focus on these fundamentals to build and/or acquire the best offensive assets.

HoIver, approximately 28% of the variation in Salary_Euro is due to differences betIen clubs. The estimated group-level variance suggests significant variability betIen clubs and their principles when it comes to player salaries. The covariance analysis indicates associations betIen certain clubs and salary for specific positions (e.g., forward, goalkeeper, midfielder). Clubs value and approach these roles differently because of their varying philosophies and styles.

Again, when comparing Ajax and FC Barcelona, the mean salary for Barcelona's defenders is significantly higher than for Ajax's defenders, indicating a potential difference in investment or player valuation betIen the two clubs when looking at defenders.

These findings provide valuable insights for making business suggestions related to player salaries, and recruitment strategies. I also hope I built an understanding of the impact of league and position dynamics on players and their compensation.

Appendix A - Variable Descriptions

Title	Category	Title In Original Data	Description
Player		Player	Player's Name
Salary_Euro	Output Variable	Annual Wages	2022 - 2023 annual salary for the individual player, this is our dependant variable
Club		NA	Name of club
Governing_Country		NA	Country where club is located
Nation_Of_Player		Nation	Nationality of the player. First, we check our records in international play at senior level. Then youth level. Then citizenship presented on wikipedia. Finally, we use their birthplace when available.
Position_1		POS	Position most commonly played by the player GK - Goalkeepers DF - Defenders MF - Midfielders FW - Forwards FB - Fullbacks LB - Left Backs RB - Right Backs CB - Center Backs DM - Defensive Midfielders CM - Central Midfielders LM - Left Midfielders RM - Right Midfielders WM - Wide Midfielders LW - Left Wingers RW - Right Wingers AM - Attacking Midfielders
Position_2		NA	Some of the players were listed with two positions, so we broke them up into two columns
Age		Age	Current age, Age at season start, Given on August 1 for winter leagues and February 1 for summer leagues.
Matches_Played		MP	Matches Played by the player or squad
Nr_Games_Started	Playing Time	Starts	Game or games started by player
Minutes		Min	Min -- Minutes
Minutes_Divided_By_90		90s	90s played, Minutes played divided by 90
Goals		Gls	Goals scored or allowed
Assists		Ast	Assists
Goals_And_Assists		G+A	Goals and Assists
Non_Penalty_Goals	Performance	G-PK	Non-Penalty Goals
Penalty_Kicks_Made		PK	Penalty Kicks Made
Penalty_Kicks_Attempted		PKatt	Penalty Kicks Attempted
Yellow_Cards		CrdY	Yellow Cards
Red_Cards		CrdR	Red Cards
Title	Category	Title In Original Data	Description
Expected_Goals		xG	Expected Goals - xG totals include penalty kicks, but do not include penalty shootouts (unless otherwise noted).
Non_Penalty_Expected_Goals	Expected	npxG	Non-Penalty xG, Non-Penalty Expected Goals
Expected_Assisted_Goals		xAG	Expected Assisted Goals - xG which follows a pass that assists a shot
Non_Penalty_Expected_Goals_Plus_Assisted_Goals		npxG+xAG -- npxG + xAG	Non-Penalty Expected Goals plus Assisted Goals - xG totals include penalty kicks, but do not include penalty shootouts (unless otherwise noted). Minimum 30 minutes played per squad game to qualify as a leader
Progressive_Carries		PrgC	Progressive Carries - Carries that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any carry into the penalty area. Excludes carries which end in the defending 50% of the pitch
Progressive_Passes	Progression	PrgP	Progressive Passes - Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch
Progressive_Passes_Received		PrgR	Progressive Passes Rec - Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch
Goalie		NA	Indicates whether or not the player is a goalie - Yes or No
Goals_Against	Goalie_Performance	GA	Goals against
Shots_On_Target_Against		SoTA	Shots on target against
Saves		Saves	Number of saves
Shots		Sh	Shots total, Does not include penalty kicks
Shots_On_Target	Shooting	SoT	Shots on target, Note: Shots on target do not include penalty kicks
Average_Shot_Distance_Yards		Dist	Average Shot Distance, Average distance, in yards, from goal of all shots taken, Minimum .395 shots per squad game to qualify as a leader, Does not include penalty kicks
Passes_Completed		Cmp	Passes Completed
Passes_Attempted	Passing	Att	Passes Attempted
Total_Passing_Distance_Yards		TotDist	Total Passing Distance, Total distance, in yards, that completed passes have traveled in any direction
Nr_of_Players_Tackled		Tkl	Tackles, Number of players tackled
Tackles_Won		TklW	Tackles Won, Tackles in which the tackler's team won possession of the ball
Blocks	Defensive_Actions	Blocks	Number of times blocking the ball by standing in its path
Shots_Blocked		Sh	Shots Blocked, Number of times blocking a shot by standing in its path
Passes_Blocked		Pass	Passes Blocked, Number of times blocking a pass by standing in its path
Touches		Touches	Touches -- Number of times a player touched the ball. Note: Receiving a pass, then dribbling, then sending a pass counts as one touch
Take_Ons_Attempted		Att	Take-Ons Attempted, Number of attempts to take on defenders while dribbling
Take_Ons_Succeded	Possession	Succ	Successful Take-Ons, Number of defenders taken on successfully, by dribbling past them, Unsuccessful take-ons include attempts where the dribbler retained possession but was unable to get past the defender
Carries		Carries	Carries, Number of times the player controlled the ball with their feet
Progressive_Carrying_Distance_Yards		PrgDist	Progressive Carrying Distance, Progressive Distance, Total distance, in yards, a player moved the ball while controlling it with their feet towards the opponent's goal