**Final Project Outline**

**Problem Description**

Using data found [online](https://fbref.com/en/) we’re 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 we can get a model that works, it would be very exciting to predict player salaries based on their statistics.

**Research Questions**

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. We 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.**

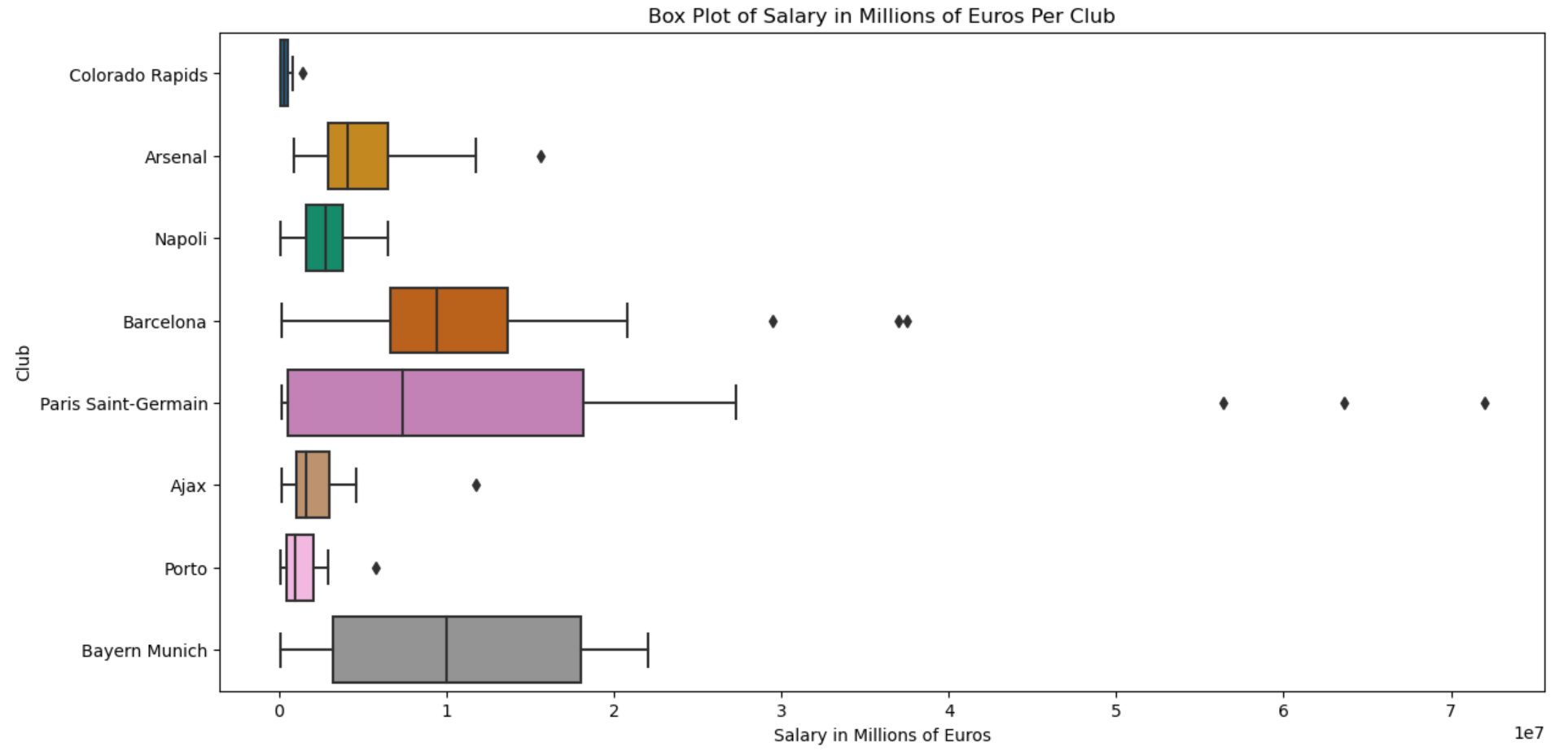
To explore our data we used Python and Tableau. We looked at outliers, top and bottom values, distributions, histograms, correlations, and descriptive statistics of our continuous and categorical variables.

**Python EDA**

[Here is a pdf link](https://drive.google.com/file/d/1ietlAW-KN2Fd2g_YNazme1yY9XpBBlhg/view?usp=sharing) to the EDA work done in Python. Below are our overall notes and three highlights from the exploration:

**EDA Notes:**

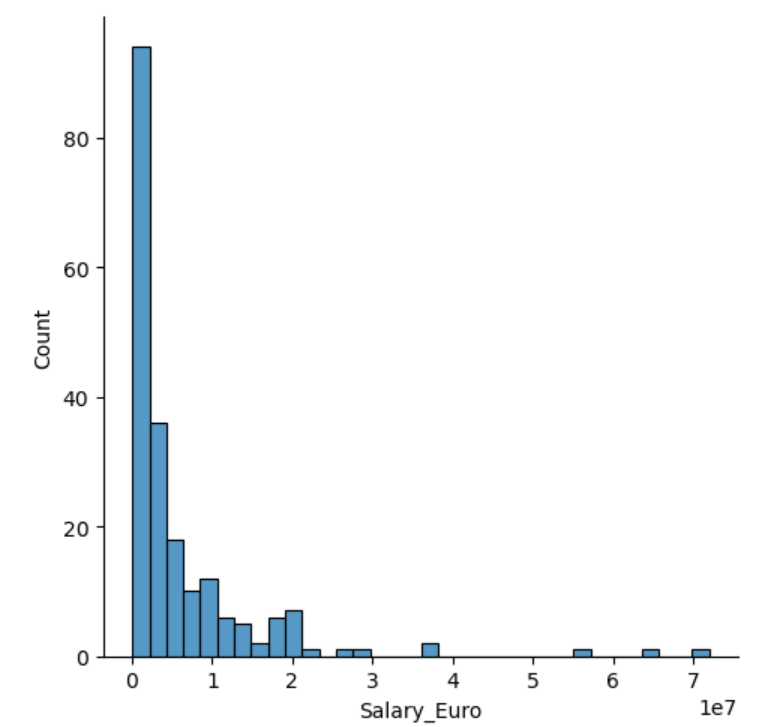
* There are 204 rows, 47 columns
* Some nulls in the column Position\_2. We can delete this column and concern ourselves with Position\_1. Alternatively, we can make it a y/n column and put a 'y' for everyone with a second position and a 'no' for people without.
* Output variable is not normally distributed.
* 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 we 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. We 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 between European teams and U.S. teams. The Colorado Rapids clearly make so much less than the EU teams.

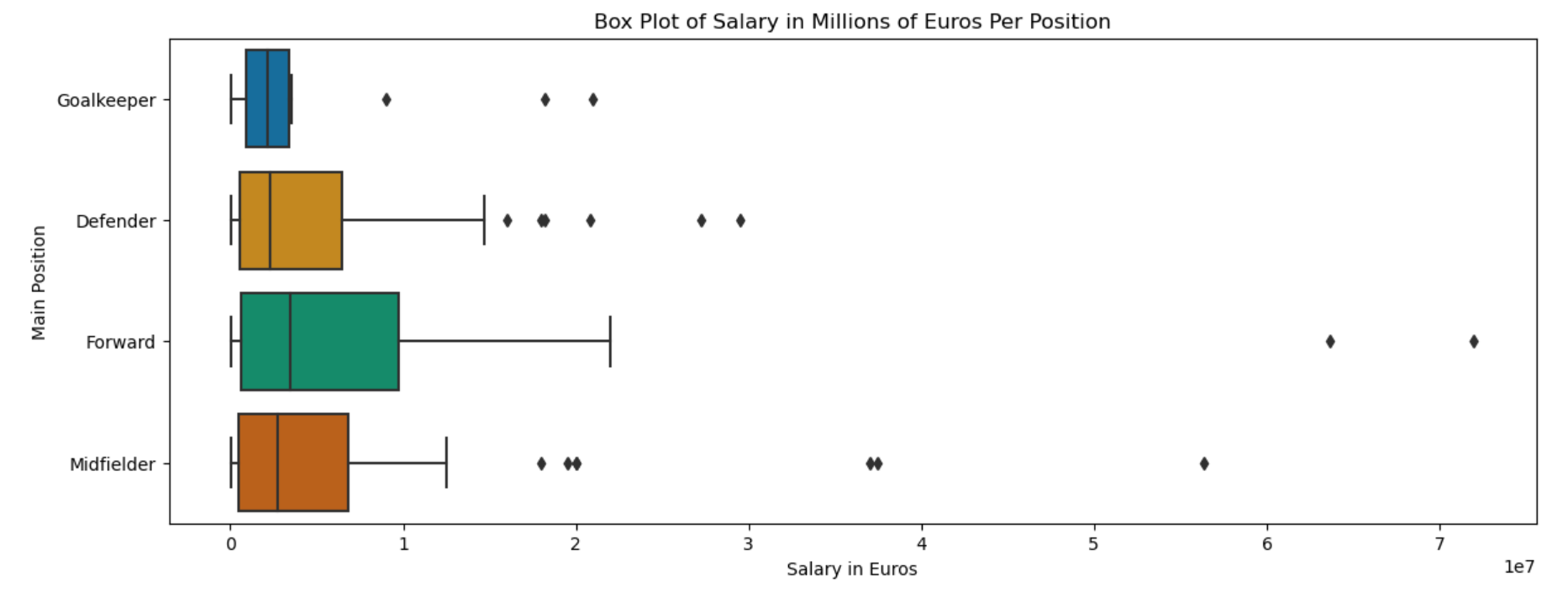
**Highlight 2:**

Histogram of Salary in Euros (Output Variable)



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

**Highlight 3:**

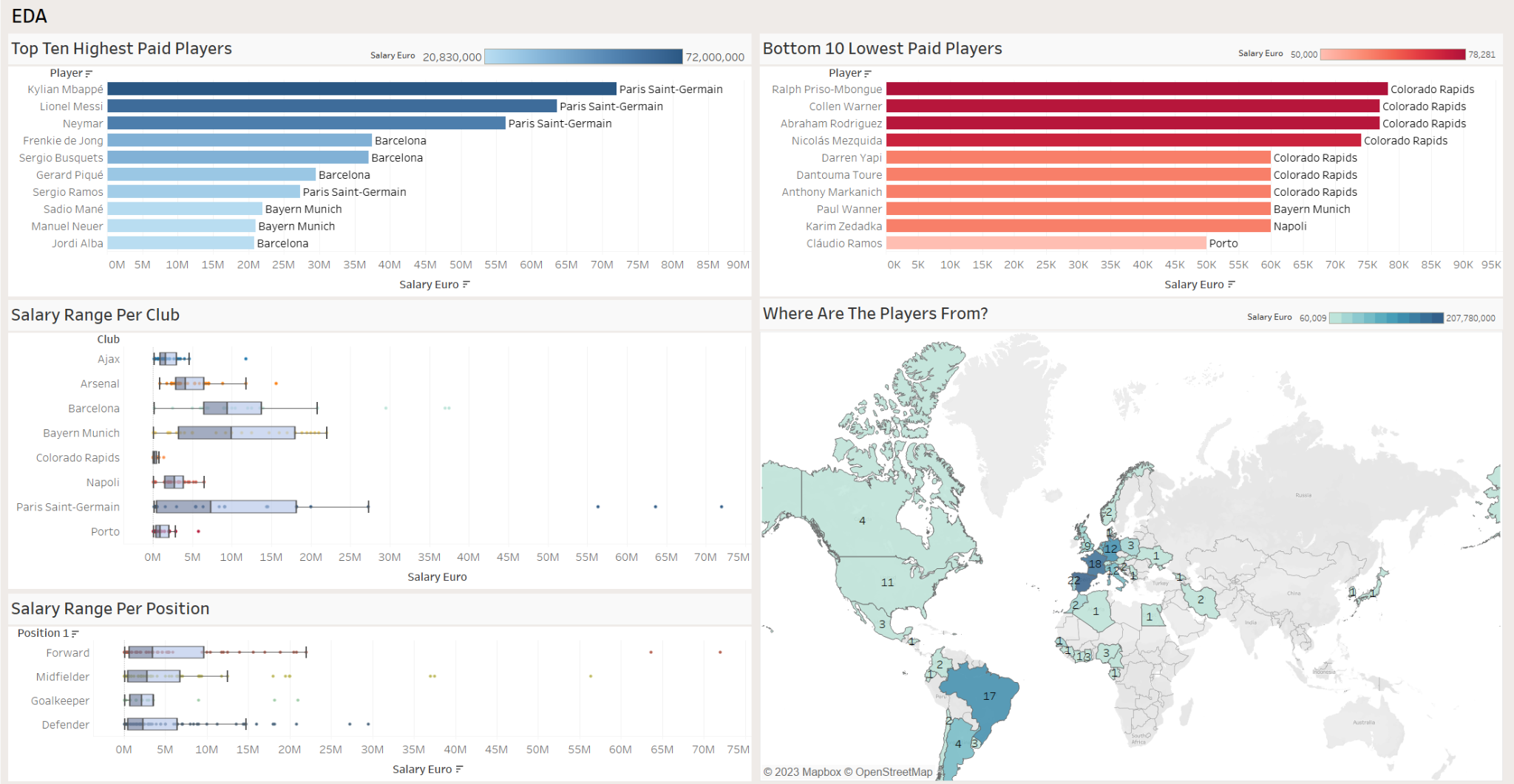
****

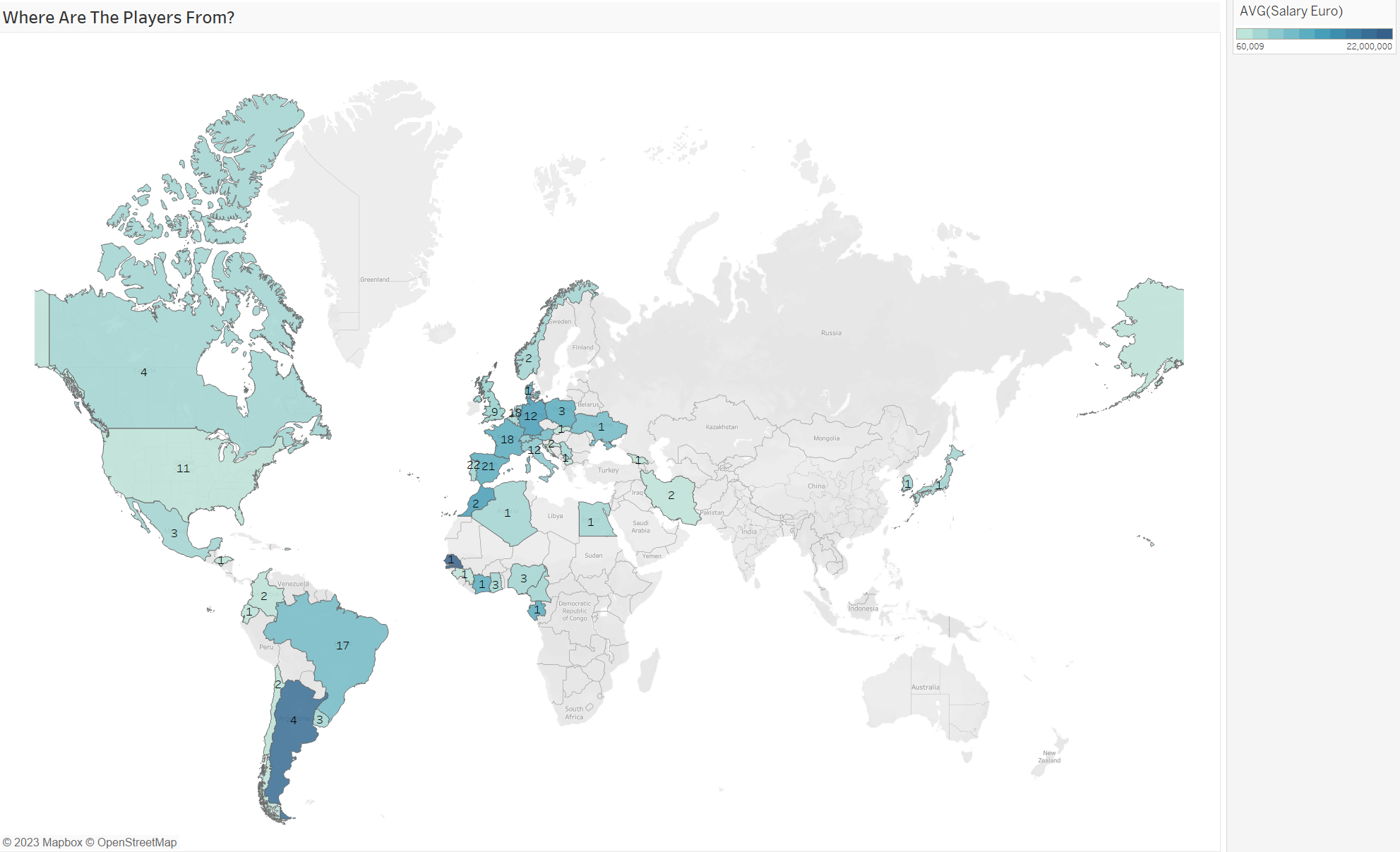
This chart makes a case for using modeling techniques to account for groups level predictions. You can see here that forwards tend to make the most money and goalies the least.

**Tableau EDA**

In Tableau we were able to quickly visualize different aspects of our dataset. We created maps, charts, and scatter plots to showcase the highlights from our EDA.

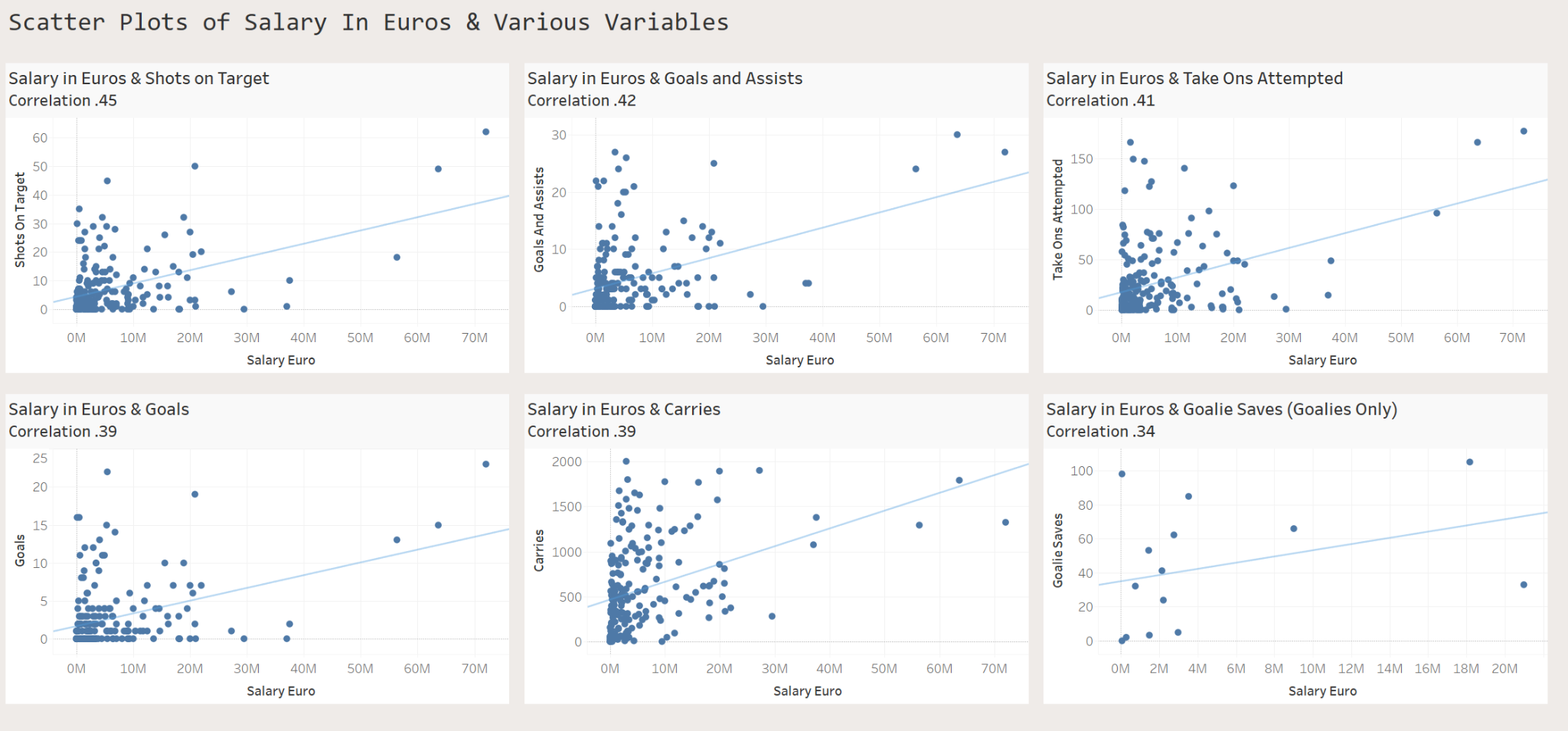
**Dashboard 1:**





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

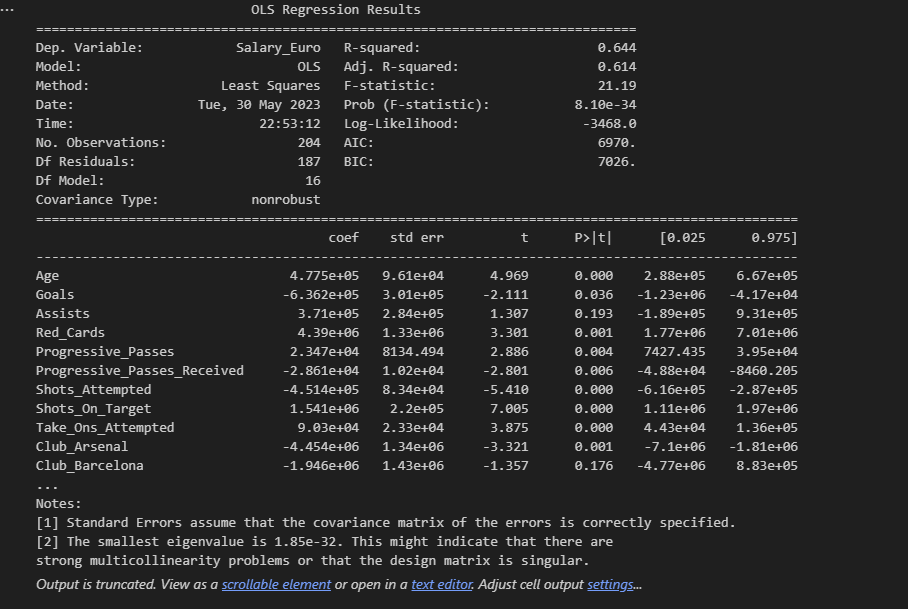
**Dashboard 2:**

****

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

**Analysis of Questions**

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

****

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 were 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.

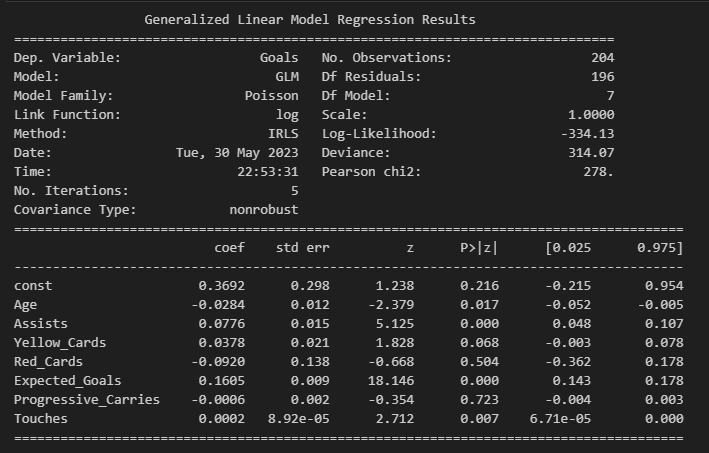
Based on these results, it is recommended to consider the statistically significant predictors in determining player salaries. Variables such as age, red cards, progressive passes, shots attempted on target, take-ons attempted, and club and governing country affiliations can be used as factors in salary 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 lower 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 lower 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 lower 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, we constructed a Poisson model to determine what factors influence goal scoring, so that scouts and clubs know what to look for.

****

**Variables used in the model:**

* **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.

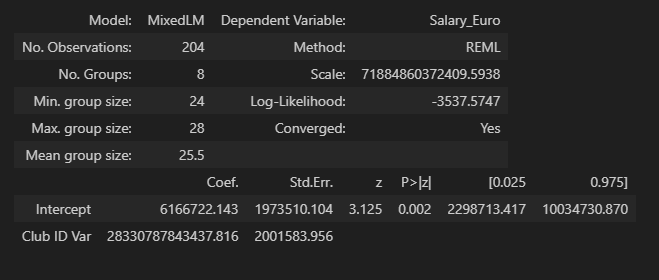
Overall, younger players, those with more assists, higher expected goals, and more touches tend to score more goals. Clubs should focus on recruiting or retaining younger players with high goal-scoring potential, such as having a high number of assists. Additionally, encouraging players to be more involved 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.

**Odd ratios for the variables in the Poisson GLM model:**

* **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?**

****

*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 between 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 between clubs regarding their impact on *Salary.*

**Covariance:**

"Club ID x C(Position\_1)[T.Forward] Cov": The covariance between the random effects of different clubs and the forward position is approximately 7.37. This suggests a moderate association between certain clubs and Salary\_Euro for forward players.

"C(Position\_1)[T.Forward] x C(Position\_1)[T.Goalkeeper] Cov": The covariance between 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 between these positions.

"Club ID x C(Position\_1)[T.Midfielder] Cov": The covariance between the random effects of different clubs and the midfielder position is approximately 11.02. This suggests a certain level of association between certain clubs and Salary\_Euro for midfielders.

"C(Position\_1)[T.Goalkeeper] x C(Position\_1)[T.Midfielder] Cov": The covariance between 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 between these positions in relation to Salary\_Euro.

These variance components and covariances provide insights into the variability and relationships between 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.

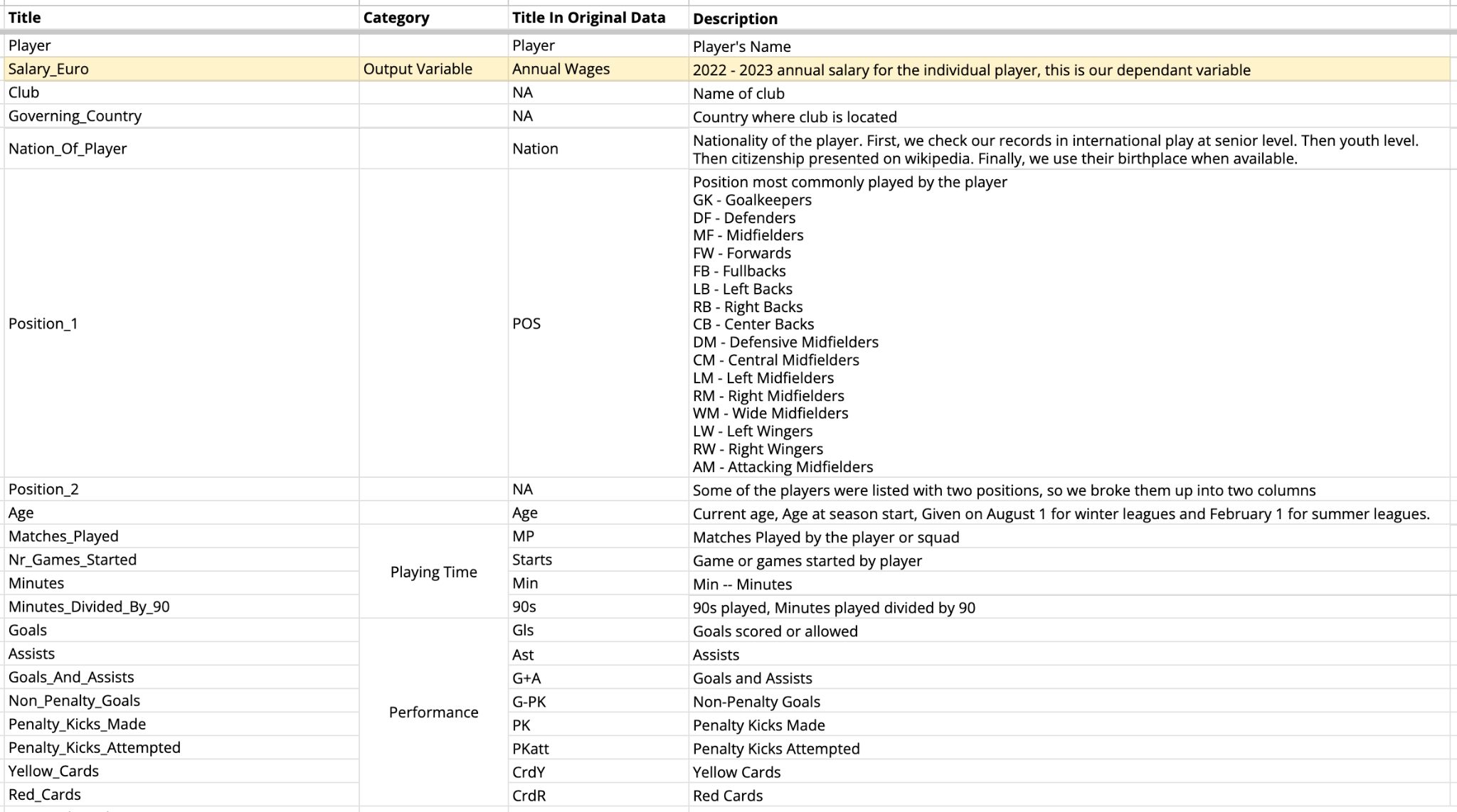
Additionally, we saw that younger players, those with more assists, and more technical skills (e.g. touches) tend to score more goals or contribute to goal scoring. Clubs should focus on these fundamentals to build and/or acquire the best offensive assets.

However, approximately 28% of the variation in Salary\_Euro is due to differences between clubs. The estimated group-level variance suggests significant variability between clubs and their principles when it comes to player salaries. The covariance analysis indicates associations between 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 between the two clubs when looking at defenders.

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

**Appendix A - Variable Descriptions**

****

****