

ADVANCED MULTIVARIATE STATISTICS PROJECT

FIFA 22 DATASET

First Name: Amr Mohamed Nazih Mohamed

Last Name: Rashad

ID: 991043

Dataset Description

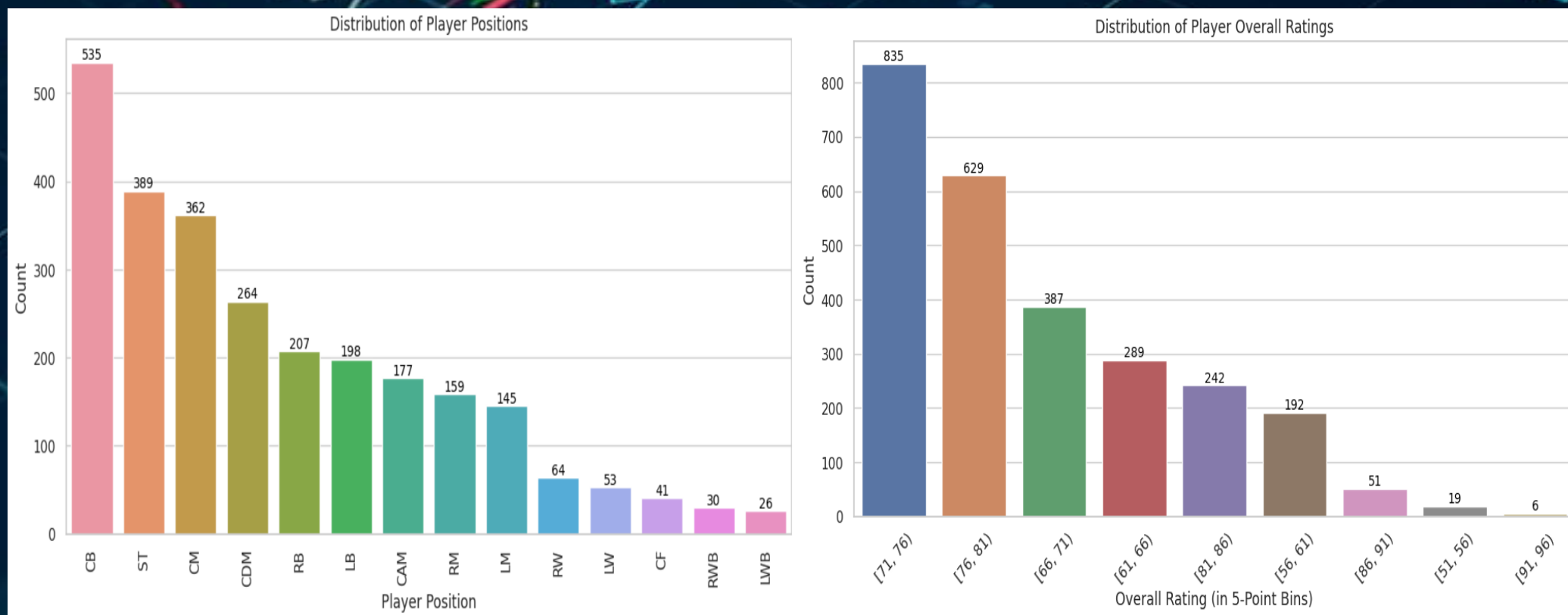
- The dataset is obtained from Kaggle, and it contains the football players data from FIFA 22.
- Originally the dataset contained 110 columns, however only the columns that are relevant to this analysis have been selected.
- Goalkeepers have been filtered out.
- Only players playing in the top 5 European leagues have been selected.
- Players' data can be described into two categories:
 - General information
 - Football attributes

Introduction

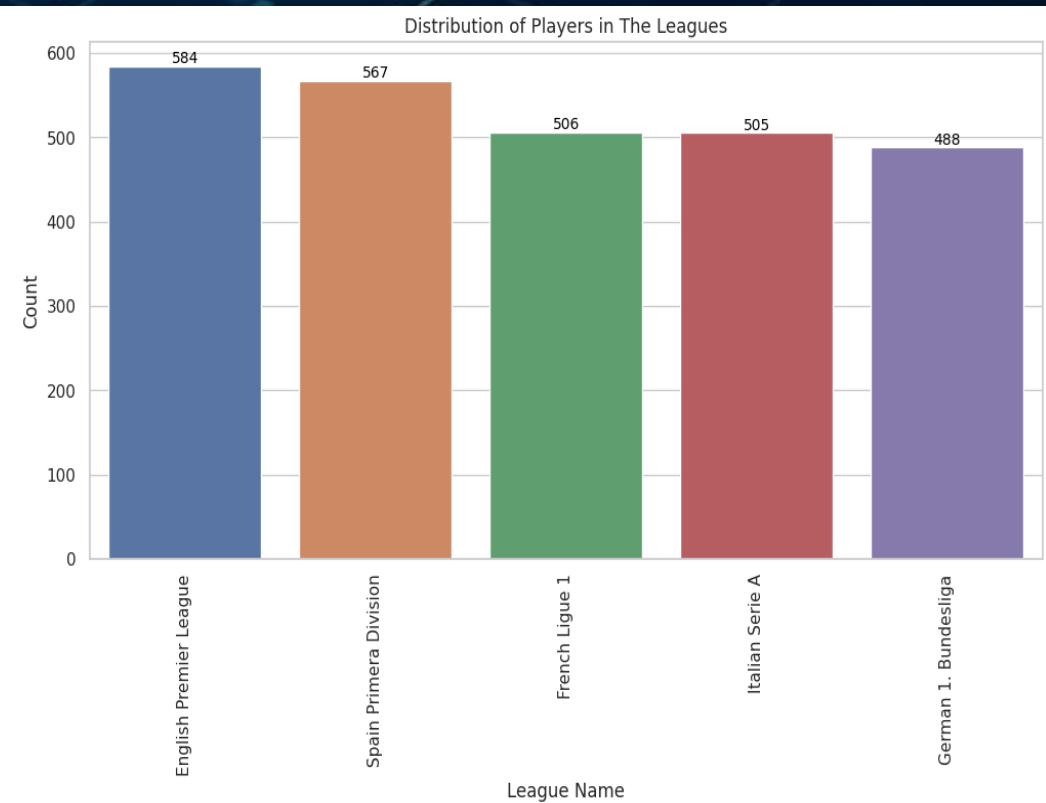
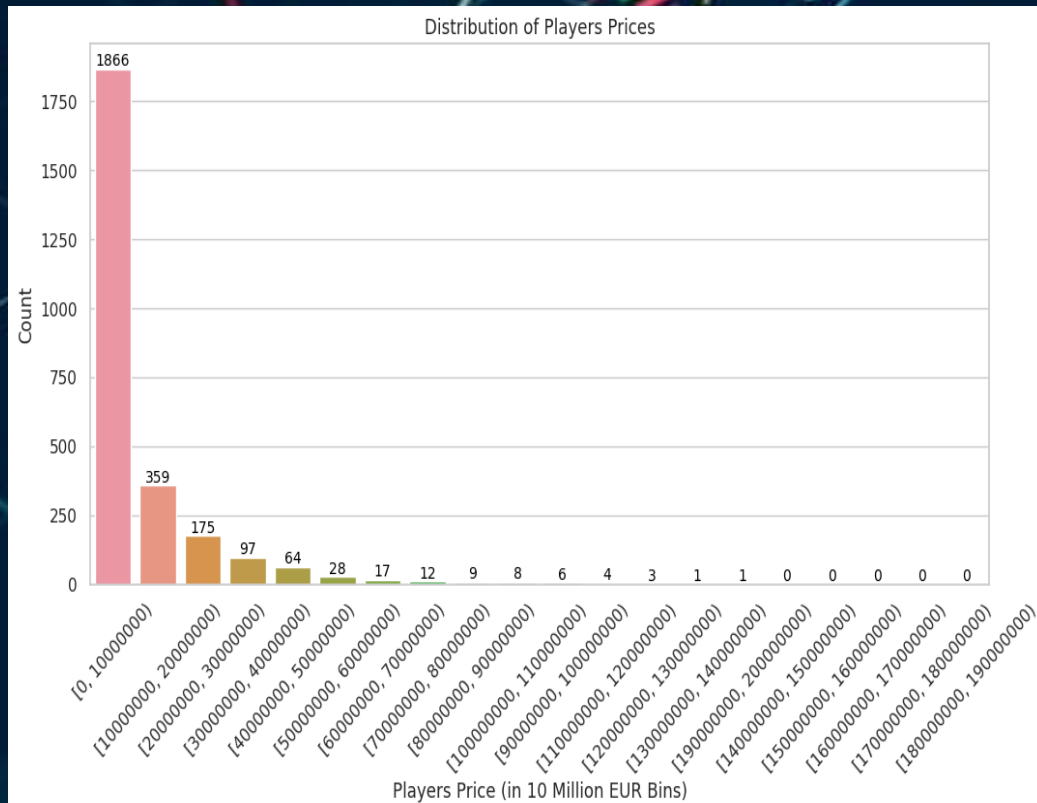
- The purpose of utilizing this dataset is to provide a straightforward and reliable dataset suitable for conducting multivariate statistical analyses.
- The types of analysis that are going to be conducted are the following:
 - Exploratory data analysis
 - Principal Component Analysis
 - Multivariate Analysis of Variance
 - Canonical Correlation Analysis

EXPLORATORY DATA ANALYSIS

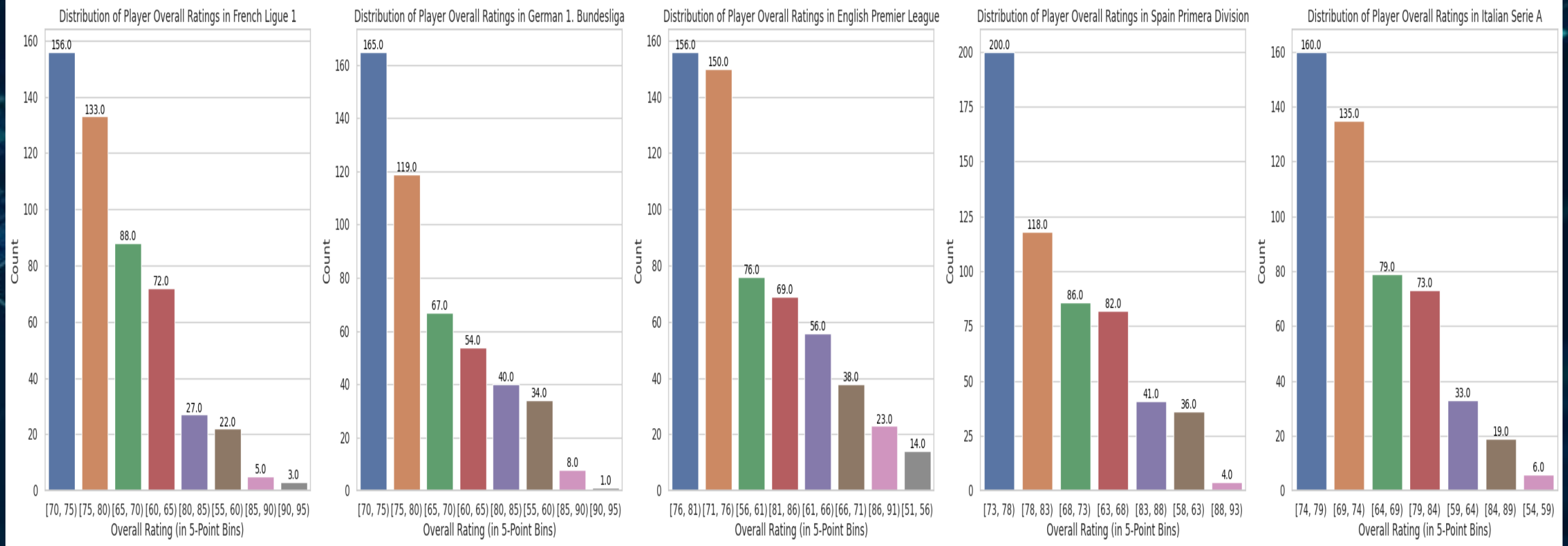
Exploratory Data Analysis



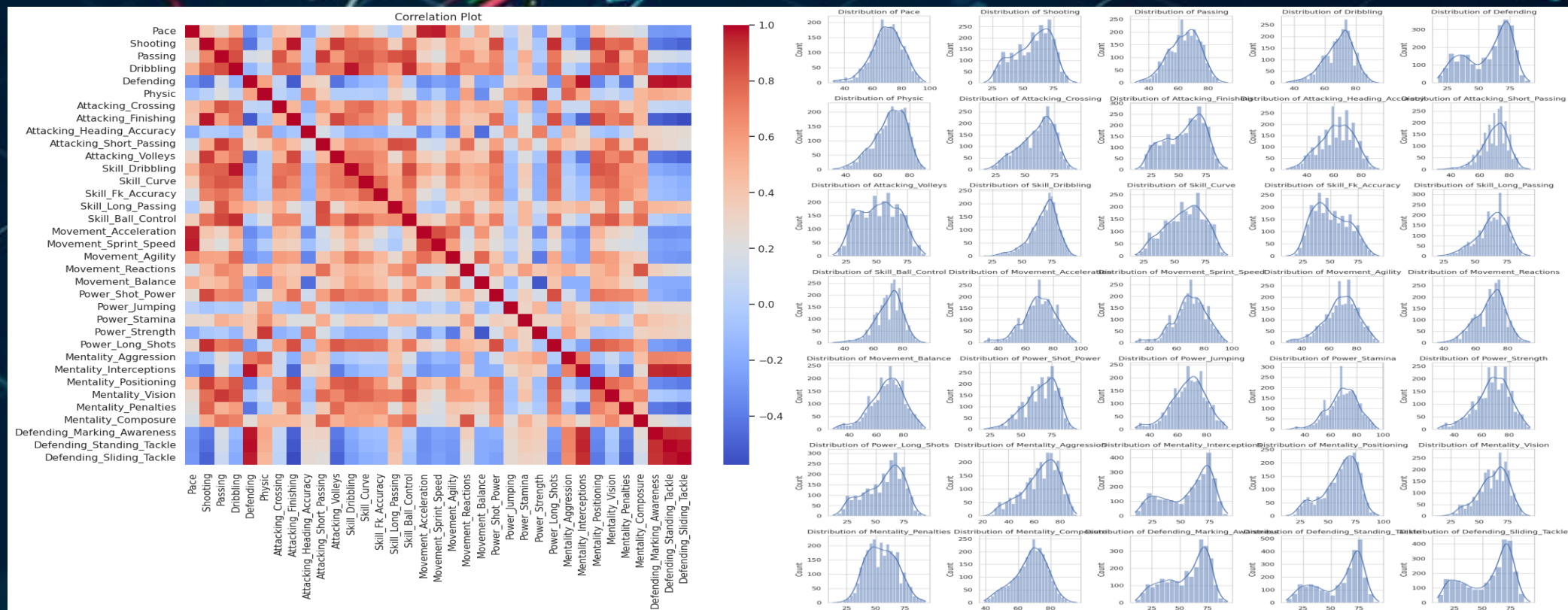
Exploratory Data Analysis



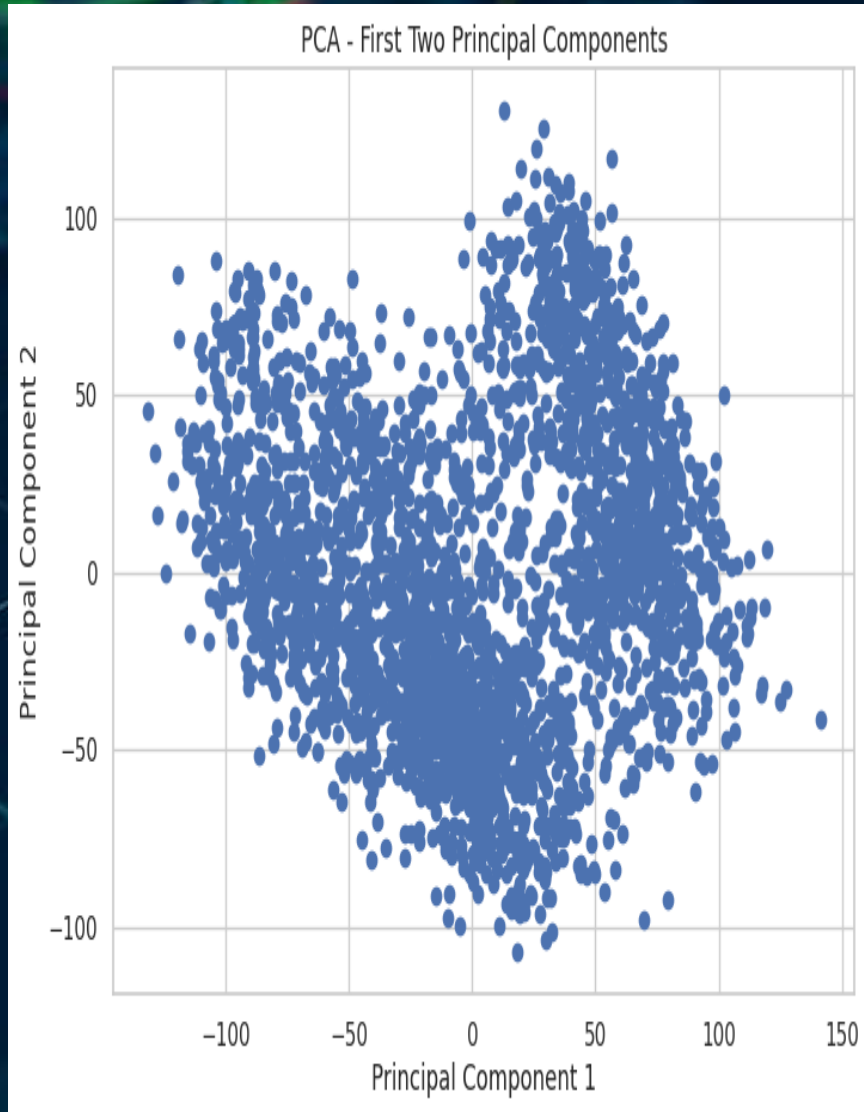
Exploratory Data Analysis



Exploratory Data Analysis



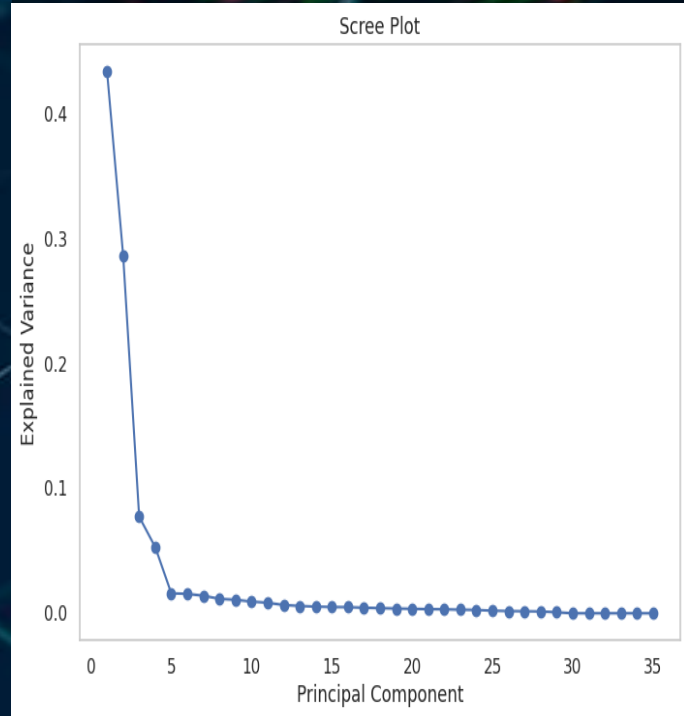
PRINCIPAL COMPONENT ANALYSIS



Principal Component Analysis

- An initial plot was created with the aim of visualizing the data in a reduced two-dimensional space.
- An assumption that can be reasonably made is that offensive and defensive players may exhibit some degree of separation or clustering in the PCA plot.
- A more graphical representation of the percentages of variation that each PC accounts for is required to determine the optimal number of dimensions that would best explain our dataset

Principal Component Analysis



Principal Component 1 explains 43.35% of the variance.
Principal Component 2 explains 28.67% of the variance.
Principal Component 3 explains 7.72% of the variance.
Principal Component 4 explains 5.30% of the variance.
Principal Component 5 explains 1.60% of the variance.
Principal Component 6 explains 1.56% of the variance.
Principal Component 7 explains 1.41% of the variance.
Principal Component 8 explains 1.16% of the variance.
Principal Component 9 explains 1.08% of the variance.
Principal Component 10 explains 0.94% of the variance.
Principal Component 11 explains 0.85% of the variance.
Principal Component 12 explains 0.66% of the variance.
Principal Component 13 explains 0.59% of the variance.
Principal Component 14 explains 0.54% of the variance.
Principal Component 15 explains 0.50% of the variance.
Principal Component 16 explains 0.49% of the variance.
Principal Component 17 explains 0.45% of the variance.
Principal Component 18 explains 0.41% of the variance.
Principal Component 19 explains 0.38% of the variance.
Principal Component 20 explains 0.36% of the variance.
Principal Component 21 explains 0.33% of the variance.
Principal Component 22 explains 0.32% of the variance.
Principal Component 23 explains 0.29% of the variance.
Principal Component 24 explains 0.27% of the variance.
Principal Component 25 explains 0.21% of the variance.
Principal Component 26 explains 0.18% of the variance.
Principal Component 27 explains 0.17% of the variance.
Principal Component 28 explains 0.15% of the variance.
Principal Component 29 explains 0.08% of the variance.
Principal Component 30 explains 0.00% of the variance.
Principal Component 31 explains 0.00% of the variance.
Principal Component 32 explains 0.00% of the variance.
Principal Component 33 explains 0.00% of the variance.
Principal Component 34 explains 0.00% of the variance.
Principal Component 35 explains 0.00% of the variance.

DataFrame with the X and Y coordinates:

	X	Y
0	141.558524	-41.621812
1	106.530757	-44.931528
2	127.396346	-32.984515
3	124.825584	-36.186879
4	79.351439	-92.259110

Correlation Coefficient: 1.4741584203520414e-16

First Component:

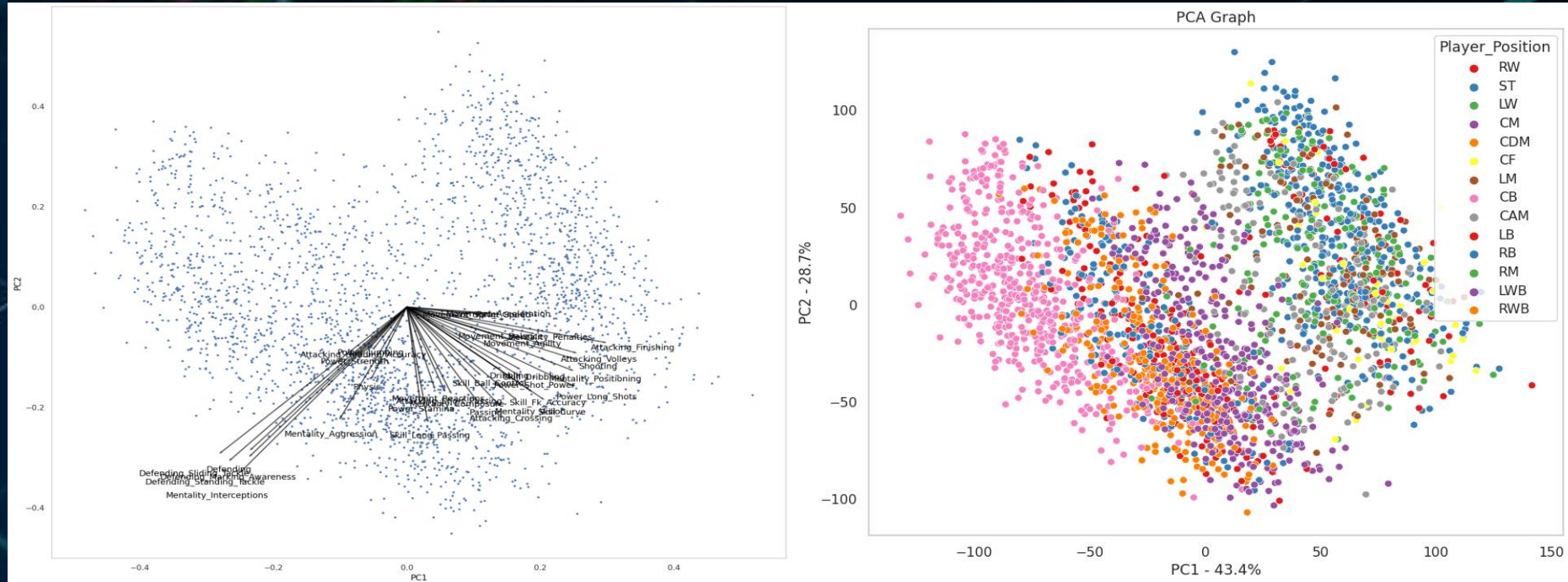
Attribute 1: Loading Score 0.1036 - Attacking_Finishing
Attribute 2: Loading Score 0.2494 - Defending_Sliding_Tackle
Attribute 3: Loading Score 0.1026 - Defending_Standing_Tackle
Attribute 4: Loading Score 0.1332 - Attacking_Volleys
Attribute 5: Loading Score -0.2318 - Shooting
Attribute 6: Loading Score -0.0516 - Mentality_Interceptions
Attribute 7: Loading Score 0.1364 - Power_Long_Shots
Attribute 8: Loading Score 0.2943 - Mentality_Positioning
Attribute 9: Loading Score -0.0564 - Defending_Marking_Awareness
Attribute 10: Loading Score 0.0554 - Defending

Second Component:

Attribute 1: Loading Score -0.0128 - Mentality_Interceptions
Attribute 2: Loading Score -0.1029 - Defending_Standing_Tackle
Attribute 3: Loading Score -0.1828 - Defending_Marking_Awareness
Attribute 4: Loading Score -0.1196 - Defending_Sliding_Tackle
Attribute 5: Loading Score -0.2818 - Defending
Attribute 6: Loading Score -0.1390 - Skill_Long_Passing
Attribute 7: Loading Score -0.1929 - Mentality_Aggression
Attribute 8: Loading Score -0.0699 - Attacking_Crossing
Attribute 9: Loading Score -0.0825 - Passing
Attribute 10: Loading Score -0.1634 - Skill_Curve

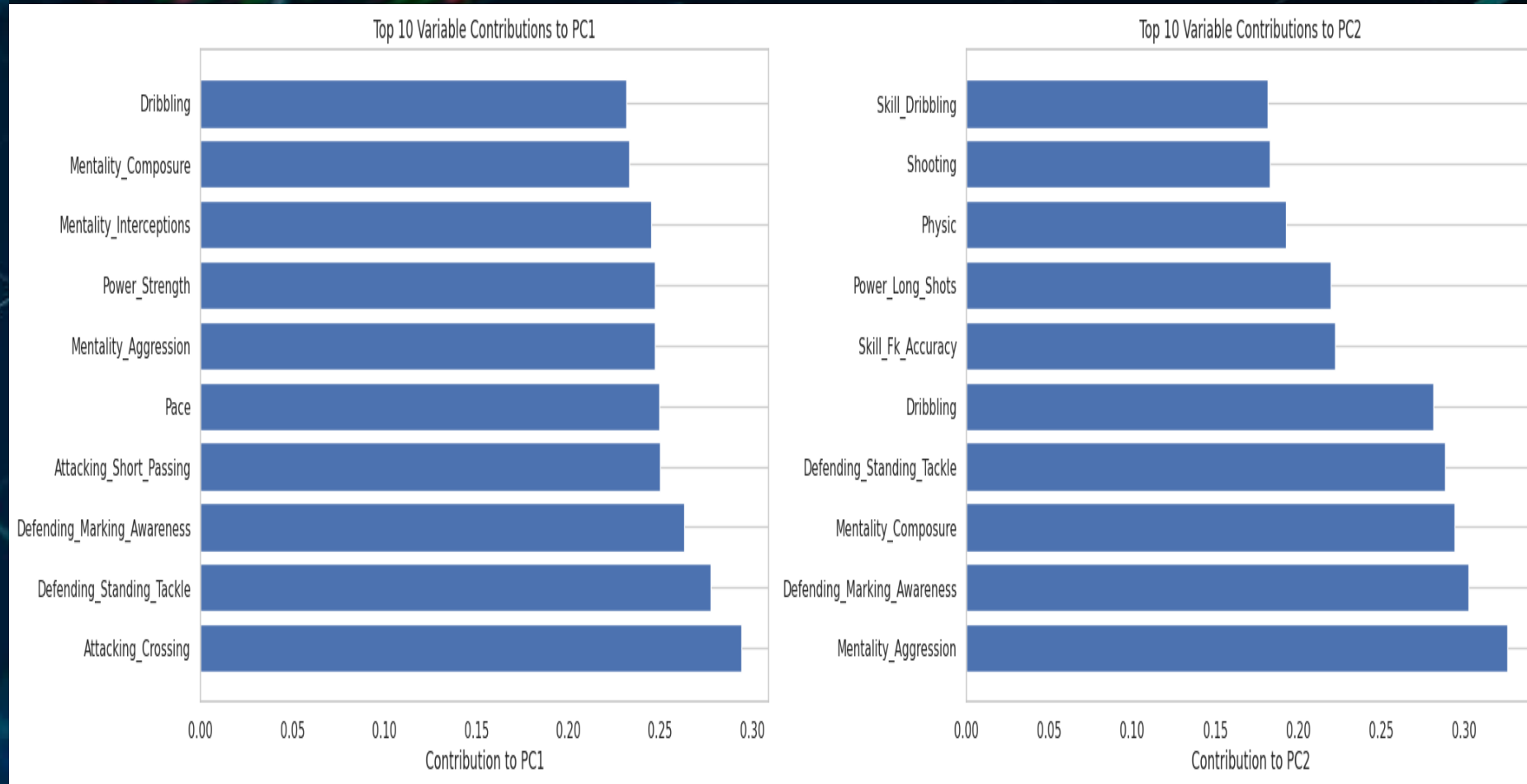
- The scree plot reveals that the first two components account for 72.02% of the variance in the original data.
- Confirming that the first two components are not correlated with each other.
- Displaying the loading scores of the top 10 attributes of each component.

Principal Component Analysis



- The arrows in the plot serve as a visual representation of the eigenvectors and their relationship to the original variables.
- There are two main directions of these arrows, lower left, and lower right. We can see that the arrows are separated into two categories, defensive attributes (lower left) and offensive attributes (lower right).
- From the scatterplot, We can clearly see that the offensive players are on the right side while the defensive players are on the left side

Principal Component Analysis



OLS Regression Results						
=====						
Dep. Variable:	Overall	R-squared:	0.678			
Model:	OLS	Adj. R-squared:	0.678			
Method:	Least Squares	F-statistic:	2789.			
Date:	Sat, 21 Oct 2023	Prob (F-statistic):	0.00			
Time:	02:39:33	Log-Likelihood:	-7475.2			
No. Observations:	2650	AIC:	1.496e+04			
Df Residuals:	2647	BIC:	1.497e+04			
Df Model:	2					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	72.3958	0.079	916.771	0.000	72.241	72.551
PC1	0.0334	0.001	23.243	0.000	0.031	0.036
PC2	-0.1254	0.002	-70.974	0.000	-0.129	-0.122
=====						
Omnibus:	19.436	Durbin-Watson:	1.329			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19.366			
Skew:	0.194	Prob(JB):	6.23e-05			
Kurtosis:	2.841	Cond. No.	55.0			
=====						

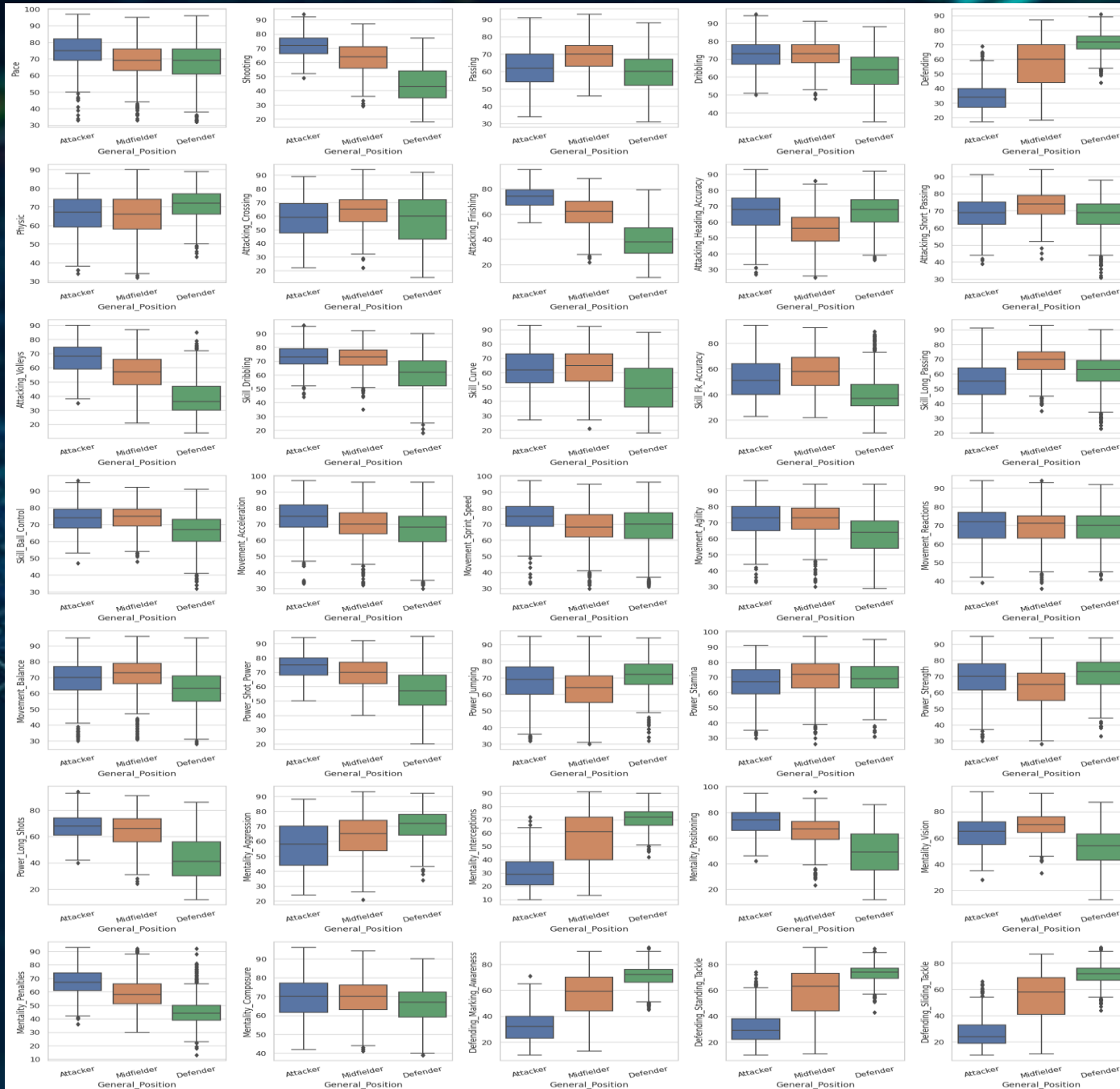
Principal Component Analysis

- It appears that the two principal components (PC1 and PC2) are significant for determining the overall rating.
- The R-squared value of 0.678 indicates that approximately 67.8% of the variance in the overall rating can be explained by the two principal components in the model.
- This suggests that PC1 and PC2 collectively have a substantial influence on the overall rating.

MULTIVARIATE ANALYSIS OF VARIANCE

Multivariate Analysis of Variance

- Testing whether player attributes vary significantly among players from different positions.
- Generalize the players' positions into three categories: Attacker, Midfielder and Defender.
- Create a set of boxplots that visualize the distribution of players' attributes based on their general positions.



Multivariate Analysis of Variance

- The MANOVA assumptions are the following:
- Groups that are compared should be independent
- Having multivariate normality
- Homogeneity of the covariance matrices
- No multicollinearity, meaning there should not be a too strong correlation between the dependent variables.
- Existence of linear relationship between the dependent variables for each group.

Multivariate Analysis of Variance

```
Shapiro-Wilk Test for "Pace": Statistic=0.9853806495666504, p-value=0.00000
Shapiro-Wilk Test for "Shooting": Statistic=0.9670935273170471, p-value=0.00000
Shapiro-Wilk Test for "Passing": Statistic=0.9895201325416565, p-value=0.00000
Shapiro-Wilk Test for "Dribbling": Statistic=0.9781668782234192, p-value=0.00000
Shapiro-Wilk Test for "Defending": Statistic=0.9231054782867432, p-value=0.00000
Shapiro-Wilk Test for "Physic": Statistic=0.9718323945999146, p-value=0.00000
Shapiro-Wilk Test for "Attacking_Crossing": Statistic=0.9686995148658752, p-value=0.00000
Shapiro-Wilk Test for "Attacking_Finishing": Statistic=0.9643842577934265, p-value=0.00000
Shapiro-Wilk Test for "Attacking_Heading_Accuracy": Statistic=0.9895935654640198, p-value=0.00000
Shapiro-Wilk Test for "Attacking_Short_Passing": Statistic=0.9734395742416382, p-value=0.00000
Shapiro-Wilk Test for "Attacking_Volleys": Statistic=0.9802194833755493, p-value=0.00000
Shapiro-Wilk Test for "Skill_Dribbling": Statistic=0.9550522565841675, p-value=0.00000
Shapiro-Wilk Test for "Skill_Curve": Statistic=0.9803358912467957, p-value=0.00000
Shapiro-Wilk Test for "Skill_Fk_Accuracy": Statistic=0.9756494164466858, p-value=0.00000
Shapiro-Wilk Test for "Skill_Long_Passing": Statistic=0.976019024848938, p-value=0.00000
Shapiro-Wilk Test for "Skill_Ball_Control": Statistic=0.9758882522583008, p-value=0.00000
Shapiro-Wilk Test for "Movement_Acceleration": Statistic=0.9836359024047852, p-value=0.00000
Shapiro-Wilk Test for "Movement_Sprint_Speed": Statistic=0.982881486415863, p-value=0.00000
Shapiro-Wilk Test for "Movement_Agility": Statistic=0.9807958602905273, p-value=0.00000
Shapiro-Wilk Test for "Movement_Reactions": Statistic=0.986078679561615, p-value=0.00000
Shapiro-Wilk Test for "Movement_Balance": Statistic=0.9817516803741455, p-value=0.00000
Shapiro-Wilk Test for "Power_Shot_Power": Statistic=0.9670257273674011, p-value=0.00000
Shapiro-Wilk Test for "Power_Jumping": Statistic=0.9840214252471924, p-value=0.00000
Shapiro-Wilk Test for "Power_Stamina": Statistic=0.9868588447570801, p-value=0.00000
Shapiro-Wilk Test for "Power_Strength": Statistic=0.9786337614059448, p-value=0.00000
Shapiro-Wilk Test for "Power_Long_Shots": Statistic=0.9604361057281494, p-value=0.00000
Shapiro-Wilk Test for "Mentality_Aggression": Statistic=0.9631208181381226, p-value=0.00000
Shapiro-Wilk Test for "Mentality_Interceptions": Statistic=0.9114976525306702, p-value=0.00000
Shapiro-Wilk Test for "Mentality_Positioning": Statistic=0.9533140659332275, p-value=0.00000
Shapiro-Wilk Test for "Mentality_Vision": Statistic=0.9763776659965515, p-value=0.00000
Shapiro-Wilk Test for "Mentality_Penalties": Statistic=0.9927334785461426, p-value=0.00000
Shapiro-Wilk Test for "Mentality_Composure": Statistic=0.9818636775016785, p-value=0.00000
Shapiro-Wilk Test for "Defending_Marking_Awareness": Statistic=0.9320263862609863, p-value=0.00000
Shapiro-Wilk Test for "Defending_Standing_Tackle": Statistic=0.8950558304786682, p-value=0.00000
Shapiro-Wilk Test for "Defending_Sliding_Tackle": Statistic=0.8984090685844421, p-value=0.00000

Confidence Intervals for "Pace":
Shapiro-Wilk Statistic: [0.98016834 0.98917739]
Shapiro-Wilk p-value: [8.92220120e-19 2.71053919e-13]

Confidence Intervals for "Shooting":
Shapiro-Wilk Statistic: [0.96090549 0.97153243]
Shapiro-Wilk p-value: [3.70597449e-26 1.51703879e-22]

Confidence Intervals for "Passing":
Shapiro-Wilk Statistic: [0.98528828 0.99208504]
Shapiro-Wilk p-value: [6.06302766e-16 7.25784086e-11]

Confidence Intervals for "Dribbling":
Shapiro-Wilk Statistic: [0.97146724 0.98292797]
Shapiro-Wilk p-value: [1.43213772e-22 2.52538330e-17]

Confidence Intervals for "Defending":
Shapiro-Wilk Statistic: [0.91545664 0.92939284]
Shapiro-Wilk p-value: [3.91286796e-36 1.18596873e-33]

Confidence Intervals for "Physic":
Shapiro-Wilk Statistic: [0.96581077 0.97687762]
Shapiro-Wilk p-value: [1.35028098e-24 2.47753196e-20]

Confidence Intervals for "Attacking_Crossing":
Shapiro-Wilk Statistic: [0.96240541 0.97353844]
Shapiro-Wilk p-value: [1.07109130e-25 9.38814447e-22]

Confidence Intervals for "Attacking_Finishing":
Shapiro-Wilk Statistic: [0.95851108 0.96927528]
Shapiro-Wilk p-value: [7.24757465e-27 2.17906323e-23]

Confidence Intervals for "Attacking_Heading_Accuracy":
Shapiro-Wilk Statistic: [0.98567376 0.99191158]
Shapiro-Wilk p-value: [1.05291602e-15 5.02897049e-11]
```

- The Shapiro-Wilk test indicates that the data in these players' football attributes is significantly deviating from a normal distribution.
- A more robust assessment of the normality of the data by taking into account the variability is introduced by bootstrapping.

Multivariate Analysis of Variance

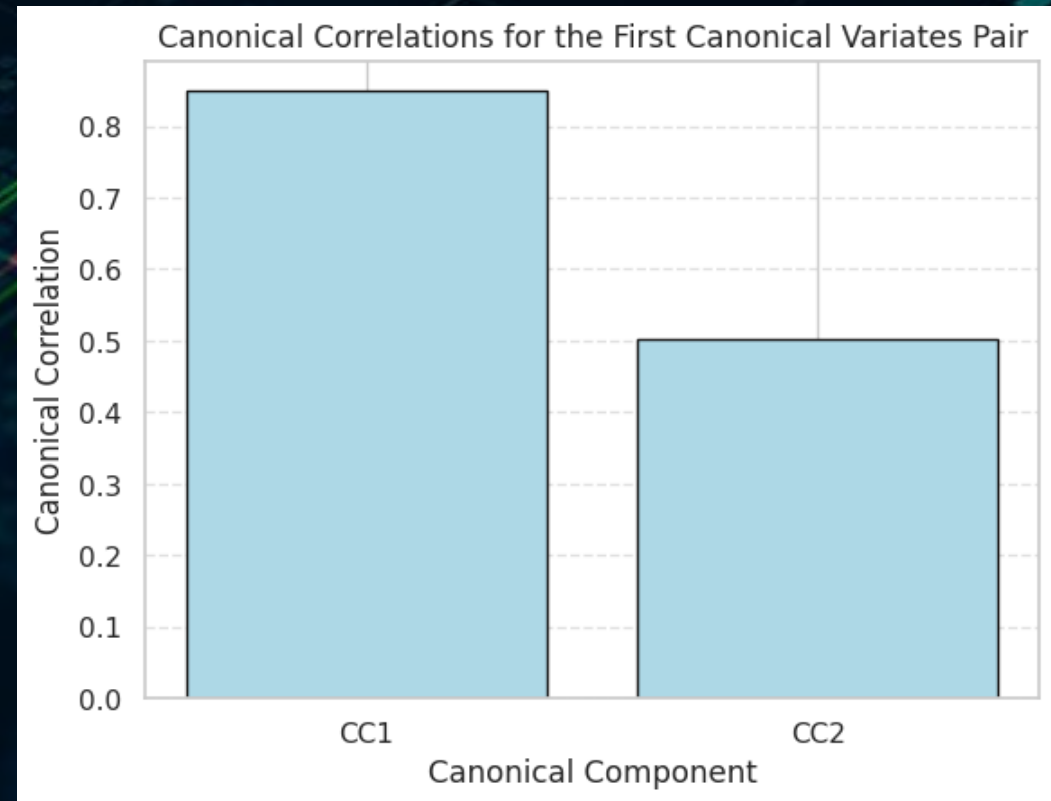
Multivariate linear model					
Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0218	35.0000	2613.0000	3343.7755	0.0000
Pillai's trace	0.9782	35.0000	2613.0000	3343.7755	0.0000
Hotelling-Lawley trace	44.7884	35.0000	2613.0000	3343.7755	0.0000
Roy's greatest root	44.7884	35.0000	2613.0000	3343.7755	0.0000
General_Position	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.1008	70.0000	5226.0000	160.5418	0.0000
Pillai's trace	1.2909	70.0000	5228.0000	135.9770	0.0000
Hotelling-Lawley trace	5.0373	70.0000	4945.6757	187.9685	0.0000
Roy's greatest root	4.0858	35.0000	2614.0000	305.1545	0.0000

- One of the tests that can be implemented that is the most robust against violations of the assumptions behind MANOVA is Pillai's Trace test.
- Null hypothesis: There are no group differences in the multivariate outcome variable(s).
- Pillai's trace value is 0.978, i.e., close to 1 & p-value is near 0 → There is a highly significant relationship between the "General Position" variable and the set of dependent variables.
- We REJECT the null hypothesis.

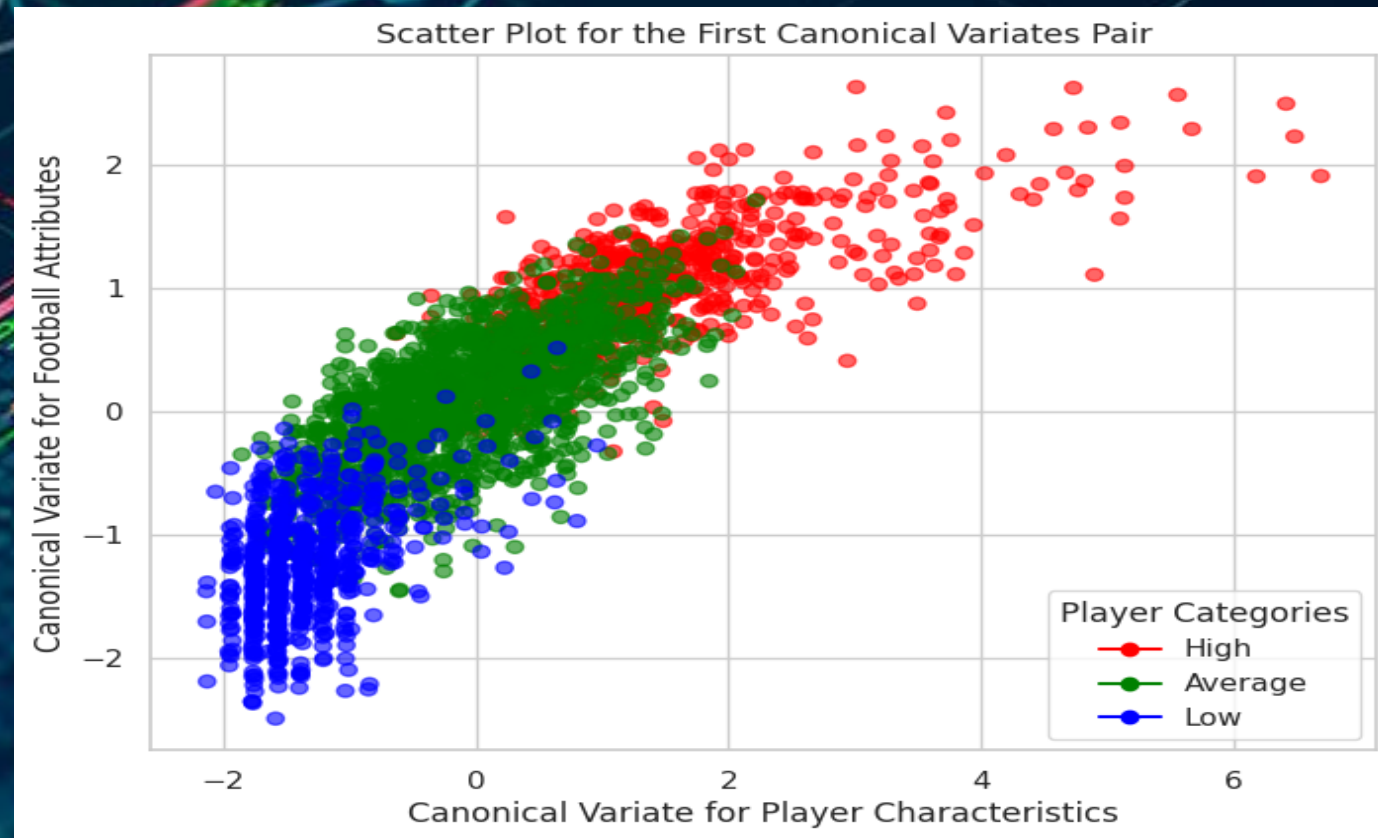
CANONICAL CORRELATION ANALYSIS

CANONICAL CORRELATION ANALYSIS

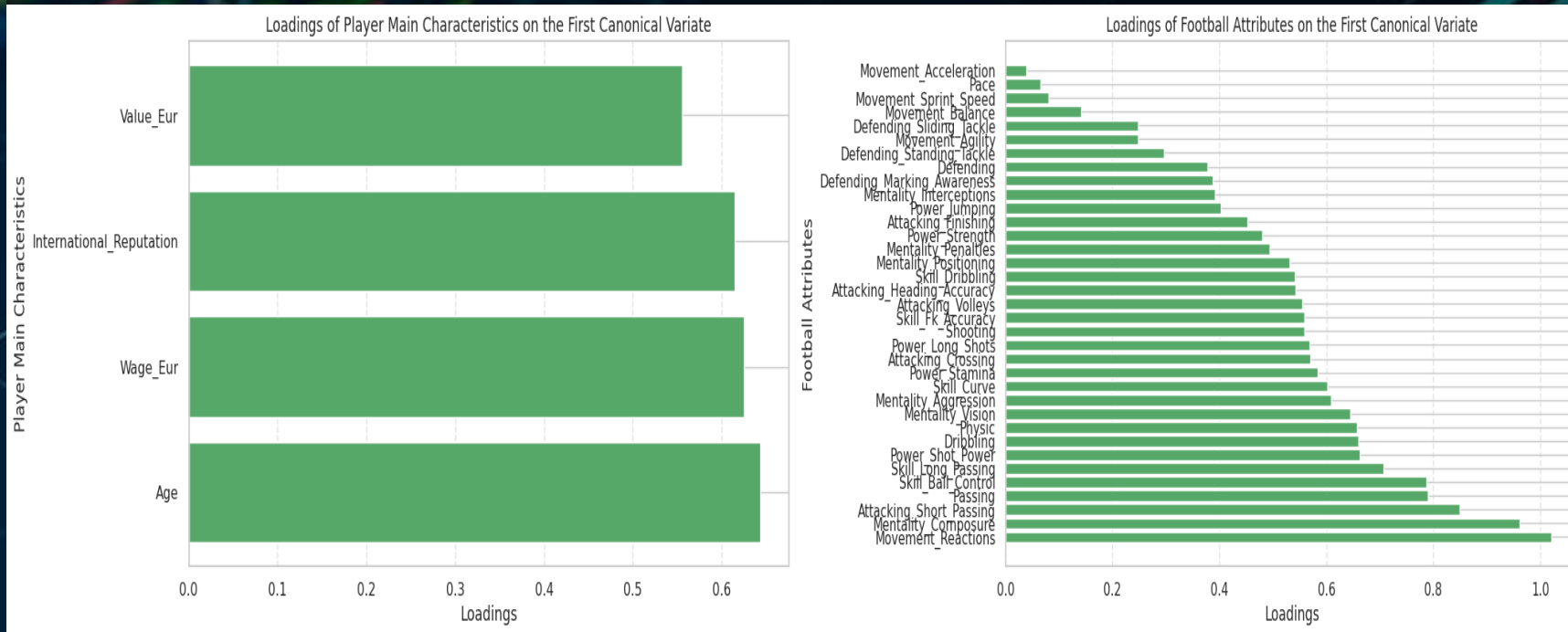
- Observe the correlation between two separate categories of data: player general information and player football attributes.
- Investigate how they collectively impact a player's overall rating.
- Generalize the players' overall rating into three categories: High, Average, Low.
- Visualizing the correlation for the first two canonical variates (CC1 and CC2).



CANONICAL CORRELATION ANALYSIS



CANONICAL CORRELATION ANALYSIS



- We do not have negative loadings.
- Magnitude of the loadings for the player's general information (left plot)
- Magnitude of the loadings for the player's football attributes (right plot)
- These loadings indicate the strength of the positive relationship of each variable with the player's overall level.

Conclusion

- This project provided valuable insights into FIFA 22 player statistics.
- It identified the key attributes influencing overall player ratings
- It examined variations in attributes across different positions
- It revealed the correlation between general player information and football attributes in determining a player's overall rating.

THANK YOU

