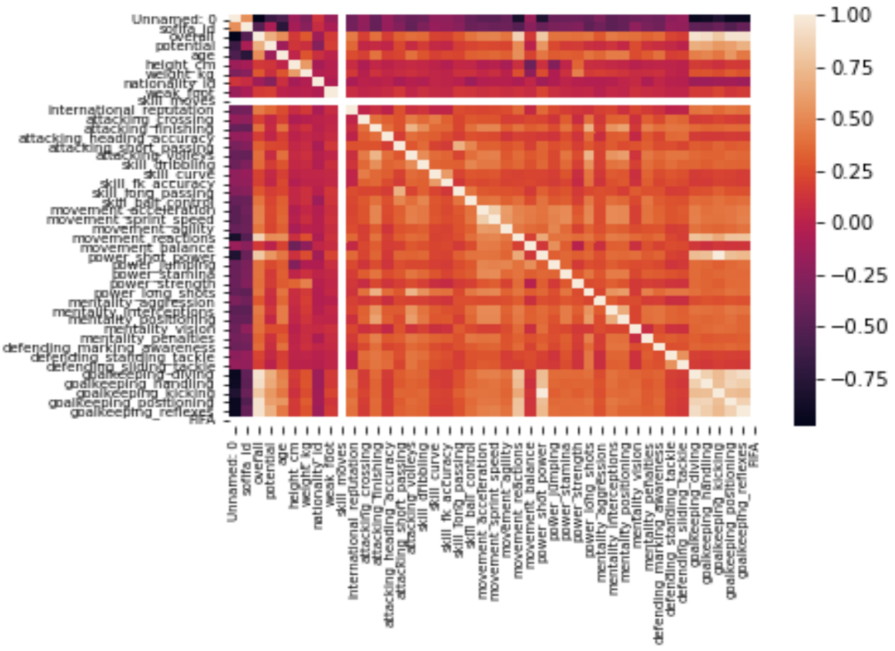
“The Beautiful Game” or more commonly referred to as football or soccer is a complex sport that involves the consideration of a mutlitude of factors in order to be succesful. With the recent adoption of technology, this sport has now been made into a successful video game franchise, FIFA. As adopting such a complex sport into a more simple and abstract concept, such as a video game, would result in the loss of its intricacy, it is interesting to see the method in which the developers have used to convert the sport into a video game. In FIFA, the main method of assessing a player’s ability is by considering their overall rating. However, in order to calculate this rating, multiple subratings and their magnitude need to be determined through elementary data analysis. These calculations are then accompanied by graphical representations as visual aid.

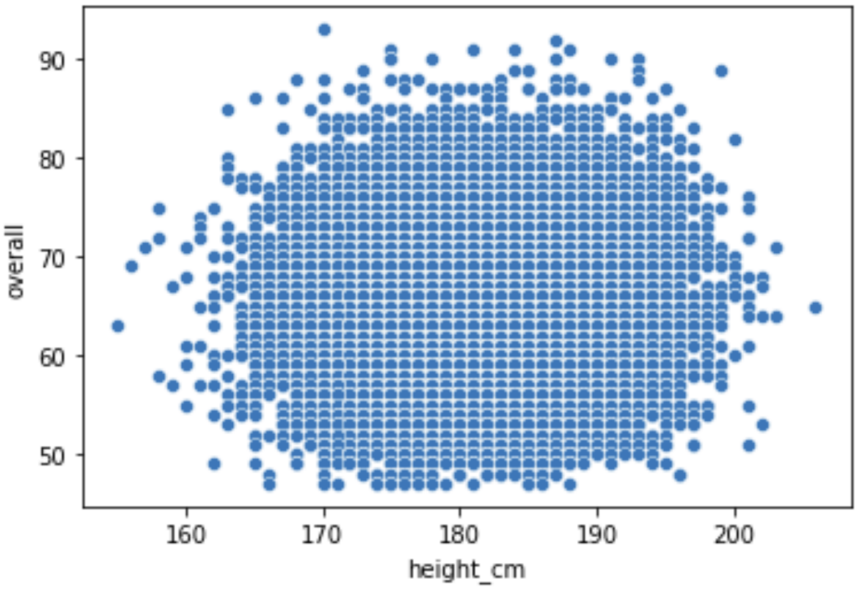
Through this experiment, the process in which a player’s rating in FIFA is determined is calculated. This process discerns which attributes are utilized for which positions and their importance. Figuring out the way in which the developers have determined the process for evaluating players can be used to determine the accuracy at which the players are evaluated and also can be used as consideration when playing the game. EDA and other forms of visualization helped describe the data in an organized manner.

EDA:
The dataset that was used in this experiment was “FIFA 22 complete player dataset” from Kaggle. There were 110 columns in this dataset in total that mainly consisted of a few introductory ones and mainly player attributes to determine their overall rating. For example, some introductory columns were their names and positions and some attributes were their long shots and composure. There were also a few columns that were completely irrelevant: unique player id on SOFIFA and URLs for flags, logos, and faces.

A heatmap visually represents the correlation of each numerical column from a lighter color, such as white to a darker color, such as black. The lighter the color, the higher the correlation and the darker the color, the lower the correlation. When the color is at its lightest, the coefficient of the correlation is 1 and when the color is at its darkest, the coefficient of the correlation is -1. This heatmap demonstrates the correlation between the columns for this experiment. Two columns that had an extremely high correlation is goalkeeping_diving and goalkeeping_reflex. On the other hand, two columns that had a n extremely low correlation is height and movement_balance.

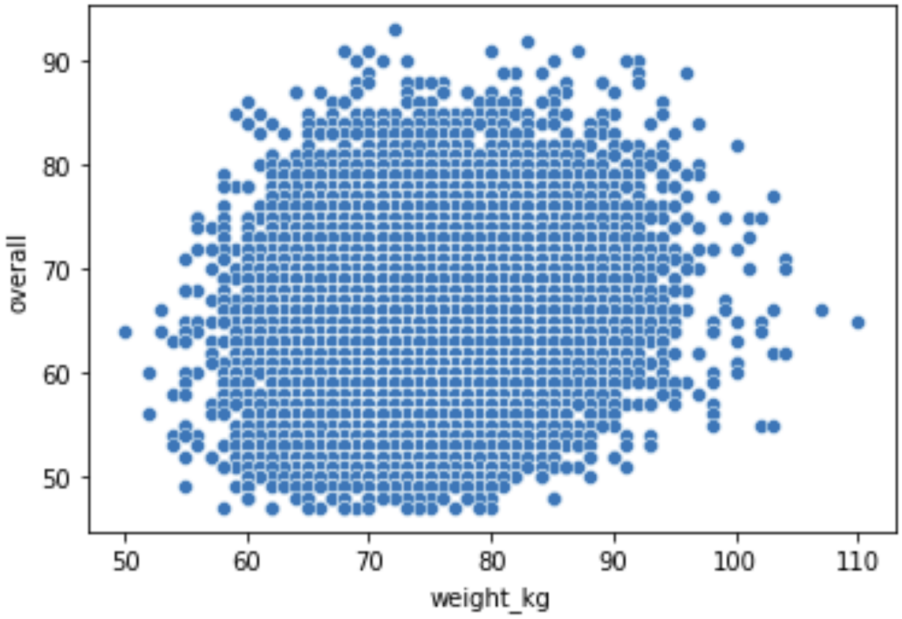


A scatterplot plots a point based on the two values from the column. Using a scatterplot helps determine the relationship between the two columns. In this scatterplot, the two columns that were used were: height_cm and overall. It seems that the points are concentrated more towards the middle of the graph and specifically towards the middle of the x-axis, around 180cm. When the height_cm disperses more towards the outer range, 160cm and 200cm, the overall rating tends to decrease in both frequency and outliers as the points closer to the average overall outnumber the points further away from the average overall.

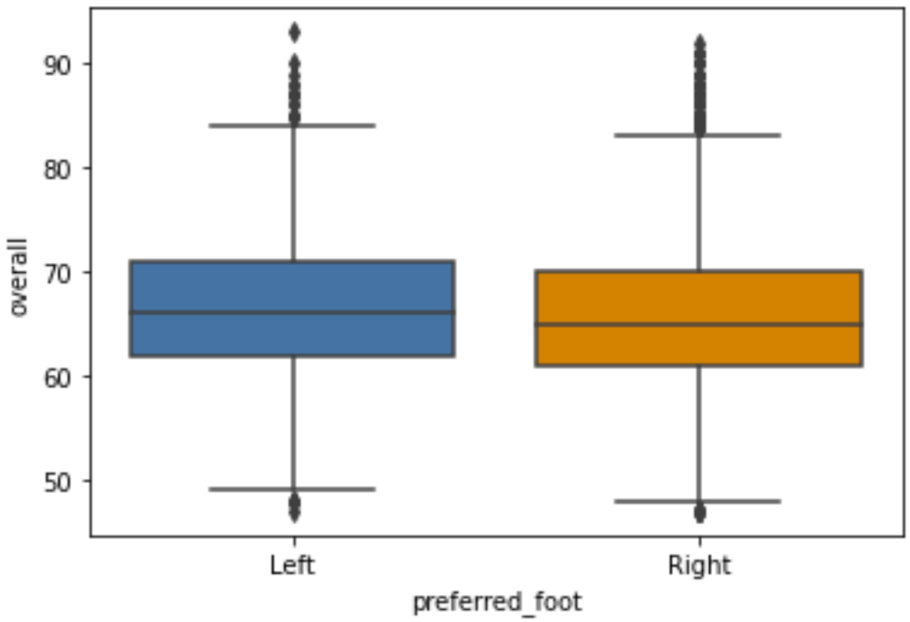


In this scatterplot, the two columns that were used were: weight_kg and overall. The points here are concentrated more towards the left of the middle of the graph, 75kg. When the graph

approaches the left of right side in terms of the x-axis, the concentration of points tends to favor one end of the y-axis. For example, when the x-axis goes towards the left, or the weight_kg decreases, the overall seems to be concentrated on the lower end. When the x-axis goes towards the right, or the weight_kg increases, the overall seems to be concentrated more on the higher end than the lower end.



A boxplot summarizes the data in a variety of ways. The line in the middle of the box represents the average, the upper line of the box represents the 75th percentile of the data, and the the lower line of the box represents the 25th percentile of the data. Although there are a much larger number of right-footed players than left-footed players, it would be predicted that the average would be higher for the right-footed players, but there was no real significant discrepancy between the two. In fact, the average of the right-footed players was lower than the average of the left-footed players.



Conclusion:
LassoCV, from sklearn.linear_model, was used to select the primary features for each position. Lasso is a linear regression model that utilizes both variable selection and regression to improve prediction accuracy and formulate a set of variables to use in a model.
These are the most relevant attributes in assessing the overall rating for a striker: 'attacking_finishing', 'attacking_heading_accuracy', 'skill_dribbling', 'skill_ball_control', 'movement_reactions', 'power_shot_power', and 'mentality_positioning'.
These are the most relevant attributes in assessing the overall rating for a center forward: 'potential', 'age', 'attacking_finishing', 'attacking_short_passing', 'skill_dribbling', 'skill_ball_control', 'movement_acceleration', 'movement_reactions', 'mentality_positioning', and 'mentality_vision'.
These are the most relevant attributes in assessing the overall rating for a winger: 'attacking_crossing', 'attacking_finishing', 'attacking_short_passing', 'skill_dribbling', 'skill_ball_control', 'movement_acceleration', 'movement_sprint_speed', 'movement_reactions', 'mentality_positioning', and 'mentality_vision'.
These are the most relevant attributes in assessing the overall rating for a wide midfielder: 'attacking_crossing', 'attacking_finishing', 'attacking_short_passing', 'skill_dribbling', 'skill_ball_control', 'movement_acceleration', 'movement_sprint_speed', 'movement_reactions', 'mentality_positioning', and 'mentality_vision'.
These are the most relevant attributes in assessing the overall rating for an attacking midfielder: 'attacking_finishing', 'attacking_short_passing', 'skill_dribbling', 'skill_ball_control', 'movement_reactions', 'mentality_positioning', and 'mentality_vision'.

These are the most relevant attributes in assessing the overall rating for a center midfielder: 'attacking_short_passing', 'skill_dribbling', 'skill_long_passing', 'skill_ball_control', 'movement_reactions', 'power_stamina', 'mentality_interceptions', 'mentality_positioning', and 'mentality_vision'.
These are the most relevant attributes in assessing the overall rating for a defensive midfielder: 'attacking_short_passing', 'skill_long_passing', 'skill_ball_control', 'movement_reactions', 'power_stamina', 'mentality_interceptions', 'defending_marking_awareness', 'defending_standing_tackle', and 'defending_sliding_tackle'.
These are the most relevant attributes in assessing the overall rating for a wing back: 'attacking_crossing', 'attacking_short_passing', 'skill_ball_control', 'movement_sprint_speed', 'movement_reactions', 'power_stamina', 'mentality_interceptions', 'defending_marking_awareness', 'defending_standing_tackle', and 'defending_sliding_tackle'.
These are the most relevant attributes in assessing the overall rating for a full back: 'attacking_crossing', 'attacking_short_passing', 'skill_ball_control', 'movement_sprint_speed', 'movement_reactions', 'power_stamina', 'mentality_interceptions', 'defending_marking_awareness', 'defending_standing_tackle', and 'defending_sliding_tackle'.
These are the most relevant attributes in assessing the overall rating for a center back: 'attacking_heading_accuracy', 'attacking_short_passing', 'movement_reactions', 'power_strength', 'mentality_aggression', 'mentality_interceptions', 'defending_marking_awareness', 'defending_standing_tackle', and 'defending_sliding_tackle'.
These are the most relevant attributes in assessing the overall rating for a goalkeeper: 'movement_reactions', 'goalkeeping_diving', 'goalkeeping_handling', 'goalkeeping_kicking', 'goalkeeping_positioning', and 'goalkeeping_reflexes'.
The attributes that Lasso deemed most relevant for each position seem fairly accurate. However, although ‘potential’ and ‘age’, for the center forward position, may be able to well predict the overall rating in general for each position, they are not really a variable that can predict the rating for each position as it is too vague. Also, there should be more defensive attributes for a full back. A full back is similar to a wing back, but a full back requires to be more active in defense while a wing back is more active in offense. The attributes that Lasso deemed most relevant to a full back seem more relevant to a wing back as they are focused more on attacking.