Info – 5709

Data Visualization

Football Player Potential Analysis

- Saigirish Suryadevara (11648660)

Introduction:

The Football Players dataset has a lot of information on the attributes and properties of football players, which are different aspects of this game. This dataset is therefore a good resource for analysts, researchers and fans who want to learn about player performance, team chemistry and tactical decisions in football context.

This project seeks to investigate the correlation between various characteristics and attributes of players as contained in the Football players dataset. In this regard, we intend to engage in extensive analysis and visualization with the aim of discovering tendencies that reveal useful information pertaining to attribute dynamics amongst soccer players.

The need to solve this problem is driven by football's complexity and sophistication. It is important to understand what factors affect a player's performance, potential and overall effectiveness in the game for different stakeholders such as coaches, scouts, and team managers. Through the application of FOOTBALL PLAYERS dataset, we can better appreciate the players attributes' dynamics and reasons behind wins within this sport.

In addition, this dataset helps us realize how vast and complicated player profiles are throughout football. By considering aspects like age, which foot they predominantly use when playing soccer matches, body stature and skill moves employed on the field; one can have an inkling why some athletes become great in their own right while others perform remarkably well as a unit.

Related Works:

Numerous precedents have researched different aspects of football player attributes and performance, providing useful findings that supplement our project aims. Here are some outstanding papers connected to our project:

- Toemen (2022) in the article “Predicting Football Player Performance Using Machine Learning Techniques”, discusses how this study uses machine learning approaches to predict performances of various players by considering their ages, positions, physical characteristics, and skill ratings. The authors conduct regression analyses to establish how the attributes of the players influence their on-field performances and so help in recruitment strategies as well as rating a player.
- Thomas, Kevin’s book (2019) “How does a football pitch impact the quality, skills and technique of footballers”: A study that looks at how a player’s preferred foot relates to his or her skill ratings in videogames related to football. Large-scale gaming datasets are analyzed by the researchers to determine how players’ choice of their favorite foot affects their performances during games, hence understanding the significance of what it is seen as a personal assumption in both real and virtual football situations.

Dataset and Attributes:

The dataset, titled ‘Football Players Data’ on Kaggle, created by Masood Ahmed, is an inclusive compendium of football players from different leagues around the globe. The dataset contains information about football players that range from demographic details to physical attributes, skill ratings to performance metrics among others. These are invaluable resources for analysts, scholars or fans who would like to get more understanding into what goes behind football player profiles and performance dynamics.

Attributes in dataset:

Although there are over 50 attributes in the dataset, we will mainly be focusing on the belove key feature from the dataset mainly in our study.

- Player Bio: This will tell you the name, age, nationality and position of each player. There is a diverse background in terms of demographic attributes which are in this data set as they represent the profiles of various footballers.
- Physical Attributes: Weight, height and body shape are just some of the physical attributes that have been recorded for these players. How well a player can perform on the field depends on such characteristics as speed, muscle power or nimbleness.
- Skill Rating: The dataset shows ratings of different parts of football technique such as shooting, dribbling, passing and defending. They are based upon performance indicators and allow players’ proficiency to be assessed quantitatively across multiple skills areas.

Methods: Python programming using Google Collab:

Google Collab can also go by the name Google Collaboratory. This is a platform that works on cloud-based and offered by Google where one can write and run Python code in an interactive browser-based system. The basis of this preference is it is:

Free GPU and TPU Access: These are Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) which are used to speed up computations, especially when training deep learning models.

No Installation Needed: When you use Google Collab, no software needs to be installed on your local machine. Just open a browsing window and connect to the internet, then begin coding.

Google Drive Integration: Collab has Google Drive built into it in such a way that users can save their notebooks and access them straight from the drive. What this means is that there is no hindrance of collaboration work as well as accessing files.

Pre-installed Libraries: There are many popular python libraries pre-installed with Google Collab including TensorFlow, PyTorch, Pandas, NumPy and Matplotlib among others.

Sharing and Collaboration: Sharing notes with other people is quite easy in Collab just like when using Google Docs. It is best suited for collaborations, code sharing and reproducibility.

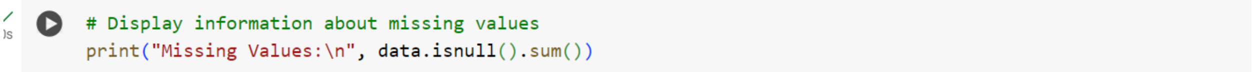
Support for Markdowns: With Collab notebooks you can make use of markdowns that will allow you to include formatted texts as well as images or even links etc. A person might want to document projects/reports or tutorials along with codes in this manner.

Python’s simplicity, versatility and extensive data manipulation and visualization support make it the best option for performing data analytics and visualization in Google Collab. By use of Python’s strong libraries, different forms of visualizations can be created like line charts, bar plots histograms scatter plot heatmaps and more to effectively analyze and communicate their insights that are driven by data.

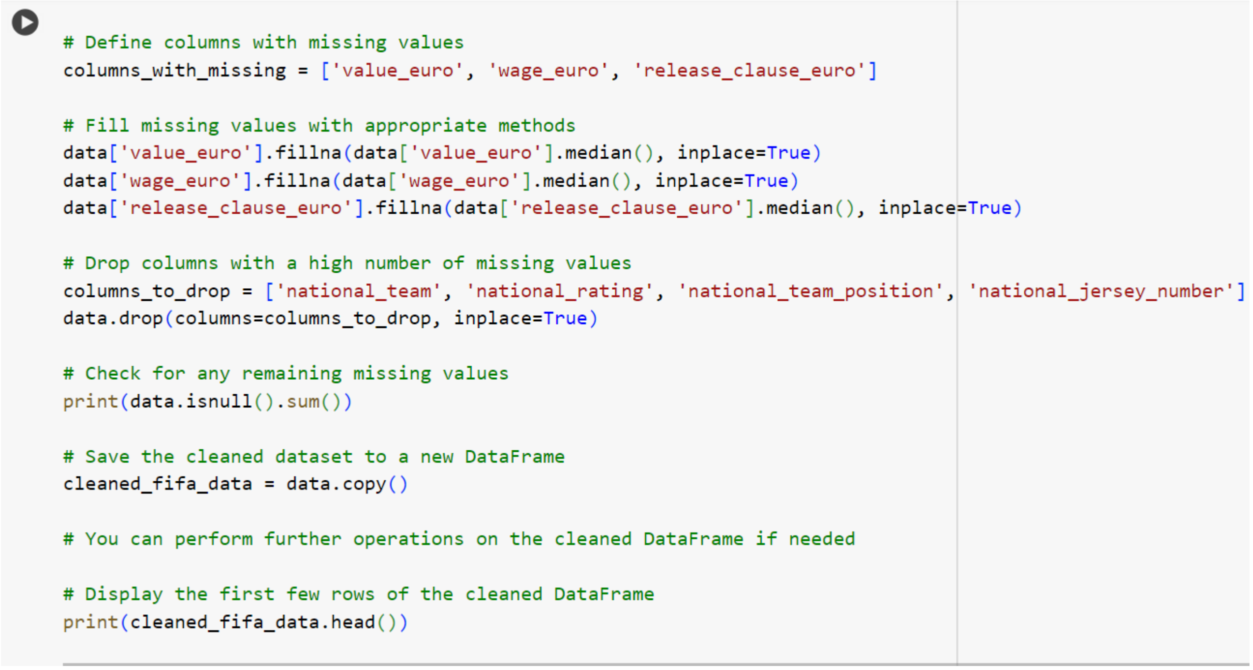
Exploratory Data Analysis:

1. Data Cleaning:

Data cleaning is an important part of the data analysis process that involves detecting and rectifying inaccuracies, inconsistencies, and missing items in the dataset to guarantee its credibility and dependability. Here’s one way to go about tidying up the Football Players dataset.



```
# Display information about missing values
print("Missing Values:\n", data.isnull().sum())
```



```
# Define columns with missing values
columns_with_missing = ['value_euro', 'wage_euro', 'release_clause_euro']

# Fill missing values with appropriate methods
data['value_euro'].fillna(data['value_euro'].median(), inplace=True)
data['wage_euro'].fillna(data['wage_euro'].median(), inplace=True)
data['release_clause_euro'].fillna(data['release_clause_euro'].median(), inplace=True)

# Drop columns with a high number of missing values
columns_to_drop = ['national_team', 'national_rating', 'national_team_position', 'national_jersey_number']
data.drop(columns=columns_to_drop, inplace=True)

# Check for any remaining missing values
print(data.isnull().sum())

# Save the cleaned dataset to a new DataFrame
cleaned_fifa_data = data.copy()

# You can perform further operations on the cleaned DataFrame if needed

# Display the first few rows of the cleaned DataFrame
print(cleaned_fifa_data.head())
```

Code 1:

This piece of code prints out information about missing values in the dataset.

It totals up the number of missing values for each column using isnull().sum() method.

Code 2:

This piece of code handles missing values as well as cleanses the dataset.

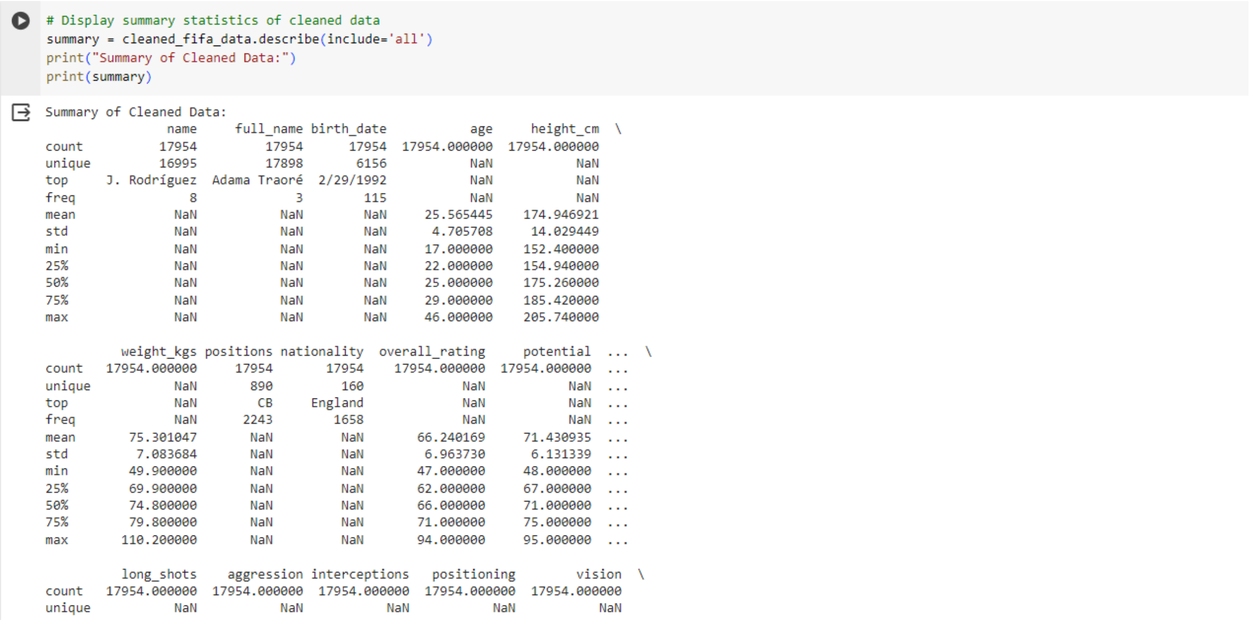
The absence of value_euro, wage_euro and release_clause_euro are features with missing in them. These missing values are replaced by median value per respective column fillna(data['column'].median()).

Those columns which have many missing values such as national_team, national_rating, national_team_position and national_jersey_number will then be removed drop()

After handling the missing data and deleting columns it checks if there are more missing data on our dataset

Finally, a cleaned version of this data set gets saved to a new Data Frame called cleaned_fifa_data using the copy() method.

2. Descriptive Analysis of cleaned data:



```python
# Display summary statistics of cleaned data
summary = cleaned_fifa_data.describe(include='all')
print("Summary of Cleaned Data:")
print(summary)
```

```
Summary of Cleaned Data:
               name     full_name birth_date           age     height_cm  \
count         17954         17954      17954  17954.000000  17954.000000
unique        16995         17898       6156           NaN           NaN
top     J. Rodríguez  Adama Traoré  2/29/1992           NaN           NaN
freq              8             3        115           NaN           NaN
mean            NaN           NaN        NaN     25.565445    174.946921
std             NaN           NaN        NaN      4.705708     14.029449
min             NaN           NaN        NaN     17.000000    152.400000
25%             NaN           NaN        NaN     22.000000    154.940000
50%             NaN           NaN        NaN     25.000000    175.260000
75%             NaN           NaN        NaN     29.000000    185.420000
max             NaN           NaN        NaN     46.000000    205.740000

          weight_kgs positions nationality  overall_rating     potential  ...  \
count   17954.000000     17954       17954    17954.000000  17954.000000  ...
unique           NaN       890         160             NaN           NaN  ...
top              NaN        CB     England             NaN           NaN  ...
freq             NaN      2243        1658             NaN           NaN  ...
mean       75.301047       NaN         NaN       66.240169     71.430935  ...
std         7.083684       NaN         NaN        6.963730      6.131339  ...
min        49.900000       NaN         NaN       47.000000     48.000000  ...
25%        69.900000       NaN         NaN       62.000000     67.000000  ...
50%        74.800000       NaN         NaN       66.000000     71.000000  ...
75%        79.800000       NaN         NaN       71.000000     75.000000  ...
max       110.200000       NaN         NaN       94.000000     95.000000  ...

          long_shots    aggression interceptions   positioning        vision  \
count   17954.000000  17954.000000  17954.000000  17954.000000  17954.000000
unique           NaN           NaN           NaN           NaN           NaN
```

The provided code snippet makes use of the describe() function to compute summary statistics for both numerical and categorical columns in the cleaned dataset. This summary shows count, mean, standard deviation, minimum, maximum and quartiles for numerical columns and count, unique values, top frequency and frequency of top value for categorical columns. Such descriptive analysis reveals central tendency, variability and categorical distribution in the data set thereby helping us understand its main features or recurring patterns that it might have.

3. Hypothesis:

Hypotheses can be developed from the available features within the FOOTBALL PLAYERS dataset.

1) Age vs Potential:

Hypothesis: Younger players might likely have higher potential ratings compared to ageing ones.

Explanation: According to this hypothesis, younger athletes who are still growing may therefore possess a better chance of improvement and hence higher potential ratings than aging sportsmen who are unlikely to improve further. This is an idea which can be used to determine whether there exists a link between age as well as prospective assessments and how it varies across different ages.

2) Preferred Foot and Skill Moves:

Hypothesis: Players who prefer to use their stronger foot may have higher skill moves ratings.

Explanation: It is believed that players who predominantly favor one foot will also have greater dexterity and technique in the other; hence, it results to raised figures in skill moves. If we can establish the link between preferred foot and skill moves rating, then we can know whether it has any correlation and how that affects a player’s capability.

3) BODY SHAPE AND PHYSICAL FEATURES:

   Hypothesis: Specific physical attributes are best demonstrated by athletes of certain body shapes.

   Explanation: This implies that diverse physical appearances such as skinny or fat may relate to physical attributes like height, weight and strength. For instance, taller players may excel in aerial duels while thinner ones may have greater pace values. As a result, establishing if there exist special body shape relationships between certain bodily characteristics of a game can be possible via the given examination.

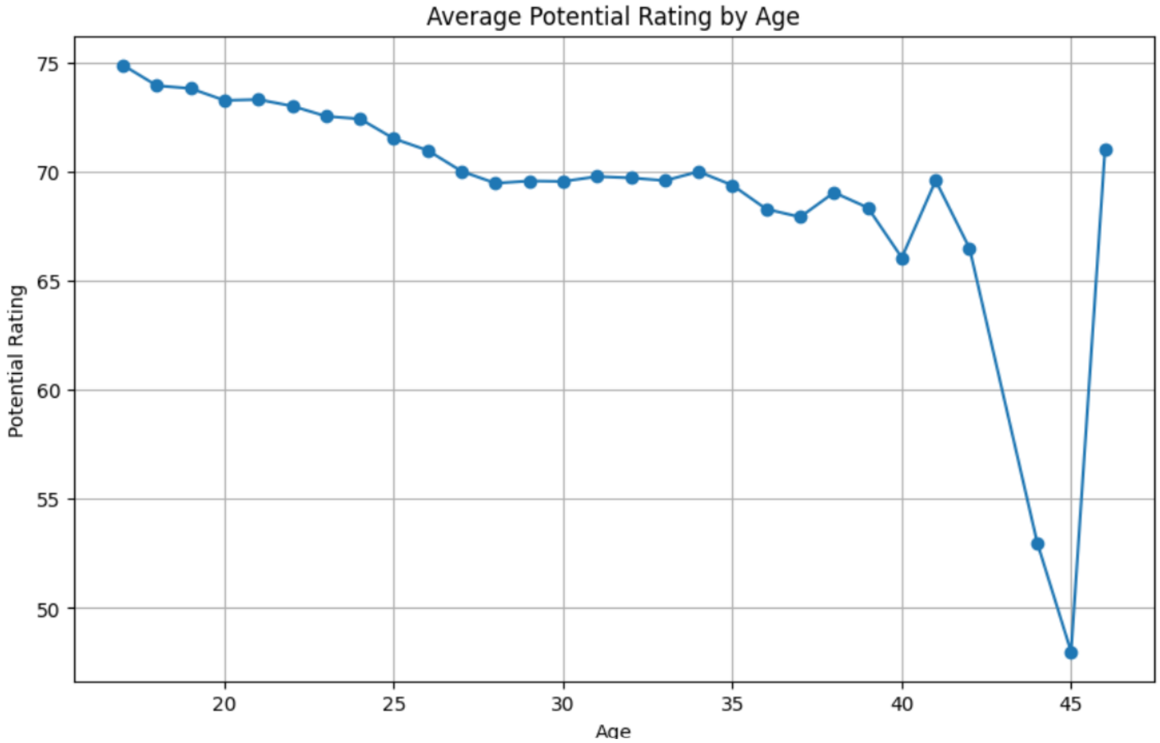
In the dataset, there are various ways to begin one's analysis, all of which entail hypothesis development that will eventually guide variable selection, visualization techniques and statistical tests aimed at exploring and validating the postulated relationships. Data visualization and hypothesis testing can reveal new patterns and enable analysts to make judgments that aid in the understanding of a data set more deeply.

Results:

1) Age vs Potential:

Hypothesis: Younger players might likely have higher potential ratings compared to ageing ones.

Line Graph:



Choice of Visualization:

A line graph is used as a data visualization, which shows the average potential rating by age. This kind of visual works well in exhibiting the connection between two quantitative variables; in this instance, age and potential rating.

Testing the Hypothesis:

This line graph shows how potential ratings change with time, thus helping to address the hypothesis that younger players have higher potential ratings than their ageing counterparts.

The graph demonstrates that at its youngest point (presumably 20 years old), the average potential rating is quite high at about 75 and gradually decreases as age increases but not in a straight line. Thus, this downplaying trend supports young players having much better promising futures when compared to old players.

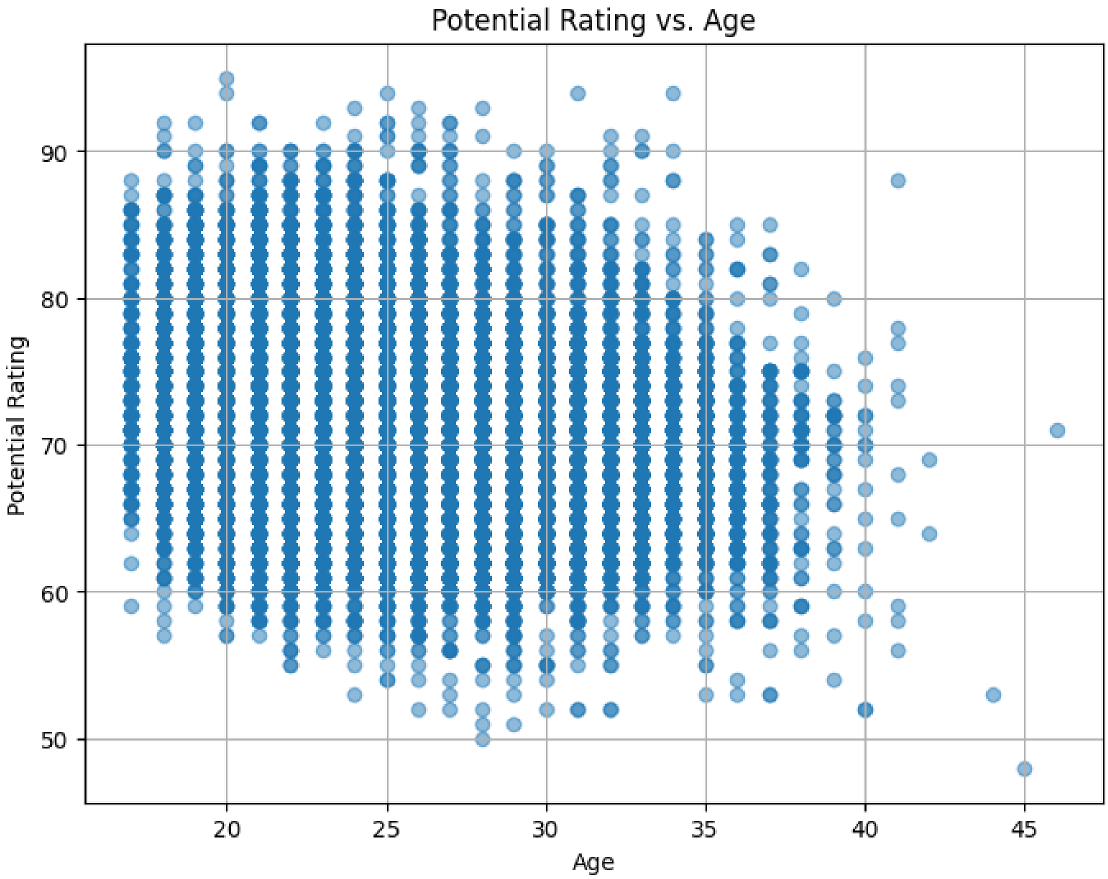
Choices of Visual Design:

Line chart choice is apt because it allows visualization of the continuous relationship between age and potential rating. The x-axis represents age while the y-axis represents the potential rating, both of which are quantitative variables.

It is one of major styles using 2D as a way of illustrating the relationship between two variables. For spatial arrangement, it has clear age on x-axis and potential rating on y-axis thereby following conventional presentation of data.

This makes the visualization easy to follow through and understand since it uses only blue color in representing both line and dots that make up data points. It does not add other colors or visual elements to distract viewers from focusing on how closely related age is with potential rating.

Scatter Plot:



Choice Of Visualization:

The visualization under discussion is in the form of a scatter plot. This is one way to display the relationship between two quantitative variables, namely potential and age. The data are represented by points.

Testing Hypotheses:

The hypothesis that younger players have higher potential ratings than older ones is supported by this scatter plot. We see a general trend where for younger ages (approximately around 20-25 years) the data points are more concentrated in the range of higher potential rating and then spread away gradually as we move down towards lower rates of potential rating with age. Despite this, there is also a wide overlap and heterogeneity in the rates of potential rating across different age groups.

Choices of Visual Design:

A scatter plot is an appropriate option because it allows for identification of individual data points, patterns and outliers and the general distribution. Age is shown on the x-axis while potential rating is at its usual place.

The 2D space design provides a standard and efficient means of illustrating the relationship between two variables. The arrangement in space has aged as it moves along the y-axis to reach potential rating.

Using two different blues for dots assist in segregating observations from each other as well as creating depth perception in visualization. Avoiding any other visual elements enables undisturbed focus on data points themselves.

Elementary Perceptual Tasks:

Referring to the scatter plot, the above elementary perceptual tasks are involved namely:

Position on a scale: The capacity to correctly locate every data point in relation to the x-axis (age) and y-axis (potential rating) scales.

Pattern detection: This refers to the proficiency of identifying and detecting patterns or trends in how data points are distributed, like that seen here where there is an evident downward trend.

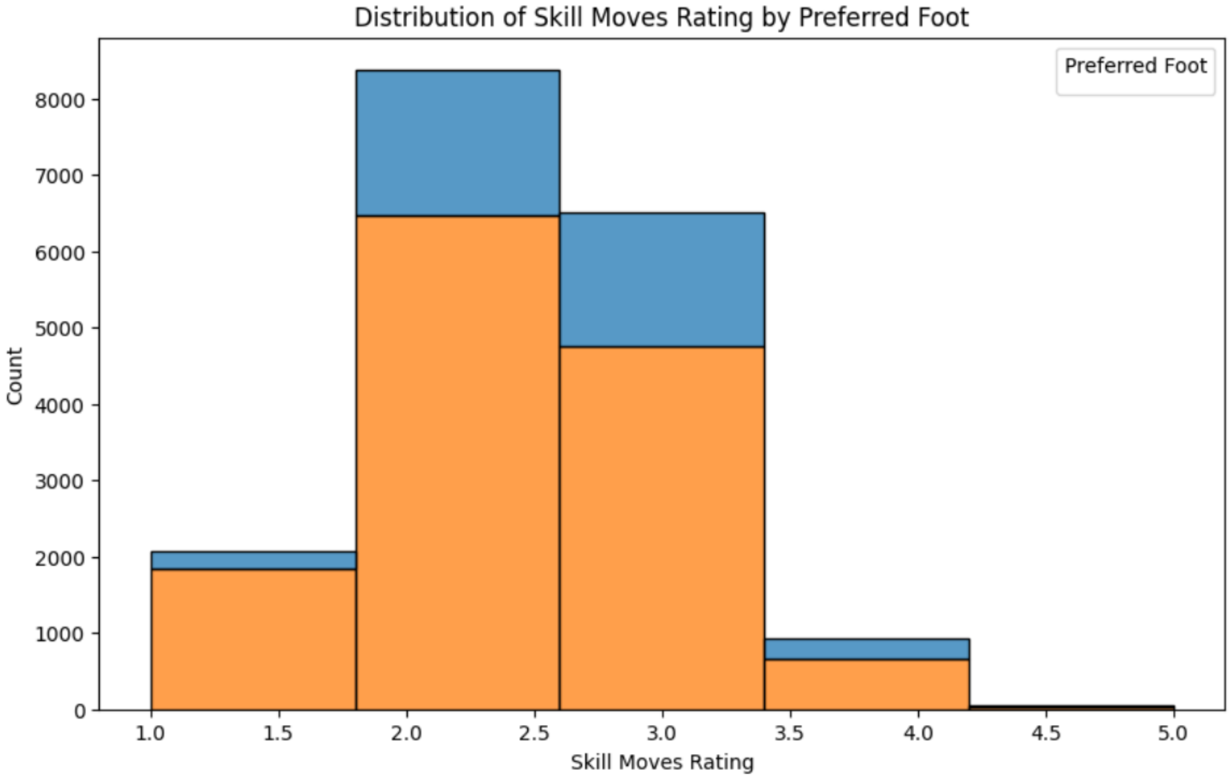
Outlier detection: Here we talk about the capacity to identify specific data that may deviate considerably from general distribution of these points that could also indicate some possible outliers.

The line chart and scatter plot together confirm the hypothesis that suggests there is a greater possibility for younger players to have better potential ratings than older ones; as shown by the former, along with other data points recorded in the latter. Nevertheless, this diagram also shows huge inconsistencies among clusters that implies it’s not just enough to consider someone’s age. Through these images, we get adequate insights on the general design and realities behind the place of age on the scale of potential rating.

2) Preferred Foot and Skill Moves:

Hypothesis: Players who prefer to use their stronger foot may have higher skill moves ratings.

Histogram:



Choice of Visualization:

The histogram explains the relationship between a player’s skill moves rating and which foot they prefer to use. This is what it has to say about the hypothesis.

Skill moves rating of 1.0 to 5.0 is represented on the x-axis, higher values show better skill moves.

On the y-axis, number or frequency of players in each skill level for all levels are indicated.

Bars are color coded with orange ones representing players who prefer their right foot and blue ones representing those who prefer their left foot.

Testing Hypothesis:

For lower skill moves ratings (1.0 – 2.5), there are significantly more blue bars (left-footed players), showing that a greater proportion of them have better skills.

The blue bars (left-footed players) were taller at higher skill move ratings (3.0 – 4.5) reflecting that many of them had good skills in this range.

This model seems to affirm our claim that players who have preference for their stronger foot do have higher skill moves ratings. There is a presumption that left-footed players, who are typically fewer in number, have improved dribbling and technical skills, possibly because they extensively favor their dominant feet.

Choice of Visual Design:

Histogram turned out to be the best visualization format since it gives an effective way of showing the distribution of a quantitative variable (skill move rating) across different categories (preferred foot). Histograms are good at visualizing and comparing frequency distributions.

Histograms are most commonly and easily understood when shown in a two-dimensional space, with the rating of skill moves on the x-axis and frequency on the y-axis. This arrangement enables a direct indication of how skill moves rating relates to player population.

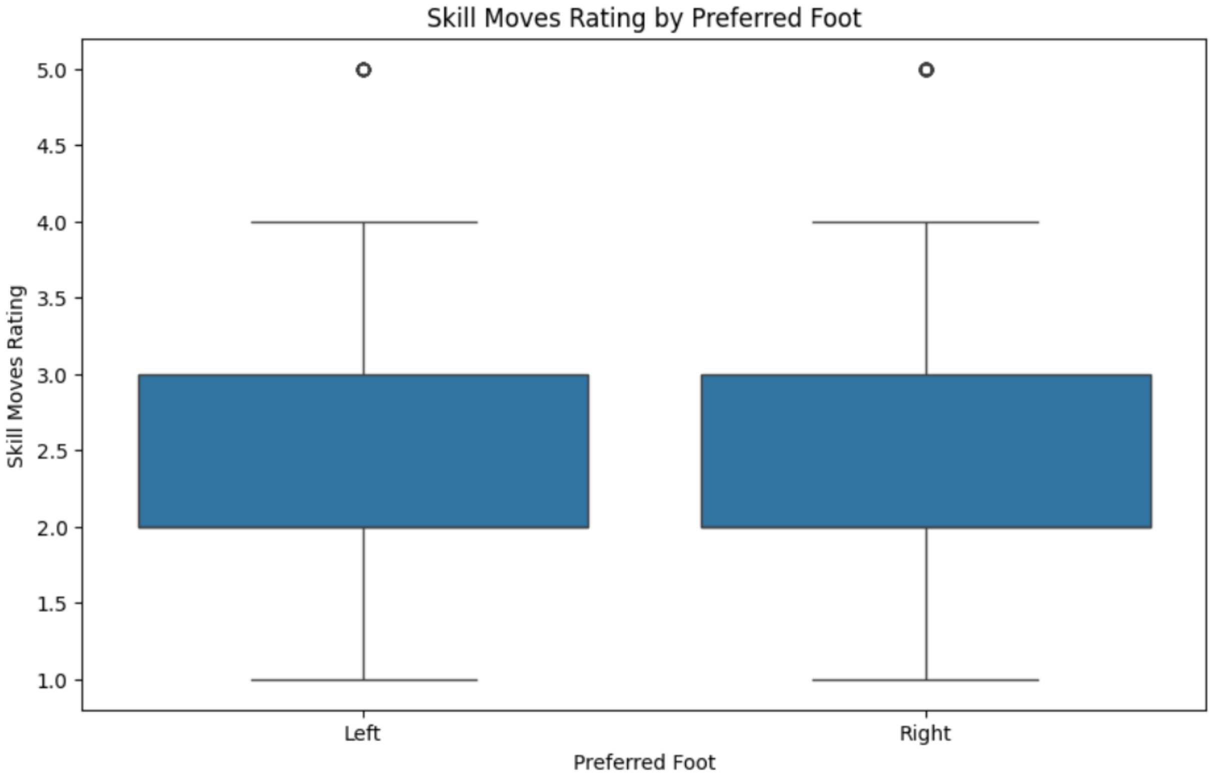
The structure is that bars for every rating of skill moves are side by side with orange bars for right-footed players adjacent to blue bars left-footed players. These groups can be easily compared all the way through the levels of skill moves ratings.

Orange color was chosen for right-footed players while left-footed players were represented in blue colors. The two colors can easily be separated visually making them distinct. Also, there is consistency in using these colors across all skill move ratings which aids in pattern recognition and interpretation.

Elementary Perceptual tasks:

While comparing heights and lengths, a histogram represents an elementary perceptual task. Basing on the frequency or count of players in each skill moves rating and preferred foot, the heights of the bars are indicated. Through comparing bar heights, distributions between two groups can be easily distinguished for differences.

Box Plot:



Choice of Visualization:

My preference for Boxplot as a graphical representation is because it helps to summarize the distribution of a numerical variable (skill moves rating) over different categories (preferred foot). Its ability to show key statistical measures like median, interquartile range and outliers in a succinct manner makes them suitable for comparing distributions.

Testing Hypothesis:

Hypothesis: Players preferring their stronger foot excel in skill moves.

Box plot evidence:

Dominant foot group has a higher median.

Dominant foot group has a larger range between the first quartile and third quartile.

Outliers for dominant foot group are above average.

This means that if these patterns are seen, it will imply that the hypothesis has been true. Nevertheless, this may need further statistical analysis to draw solid conclusions.

Choice of Visualization design:

Box plot was chosen as the appropriate visualization format since it effectively summarizes and displays the distribution of a quantitative variable (skill moves rating) across various categories (preferred foot). They represent important statistical measures such as median, interquartile range, and outliers in a concise manner which makes them suitable for comparing distributions.

The y-axis has skill moves rating while x- axis has preferred foot groups (Left and Right). This design enables easy comparisons between the two groups along the horizontal axis.

The left-footed group’s box is on the left side while that of right-footed group is located on the right of its graph. It corresponds to preferences in foot directly making it easy to understand and compare distributions.

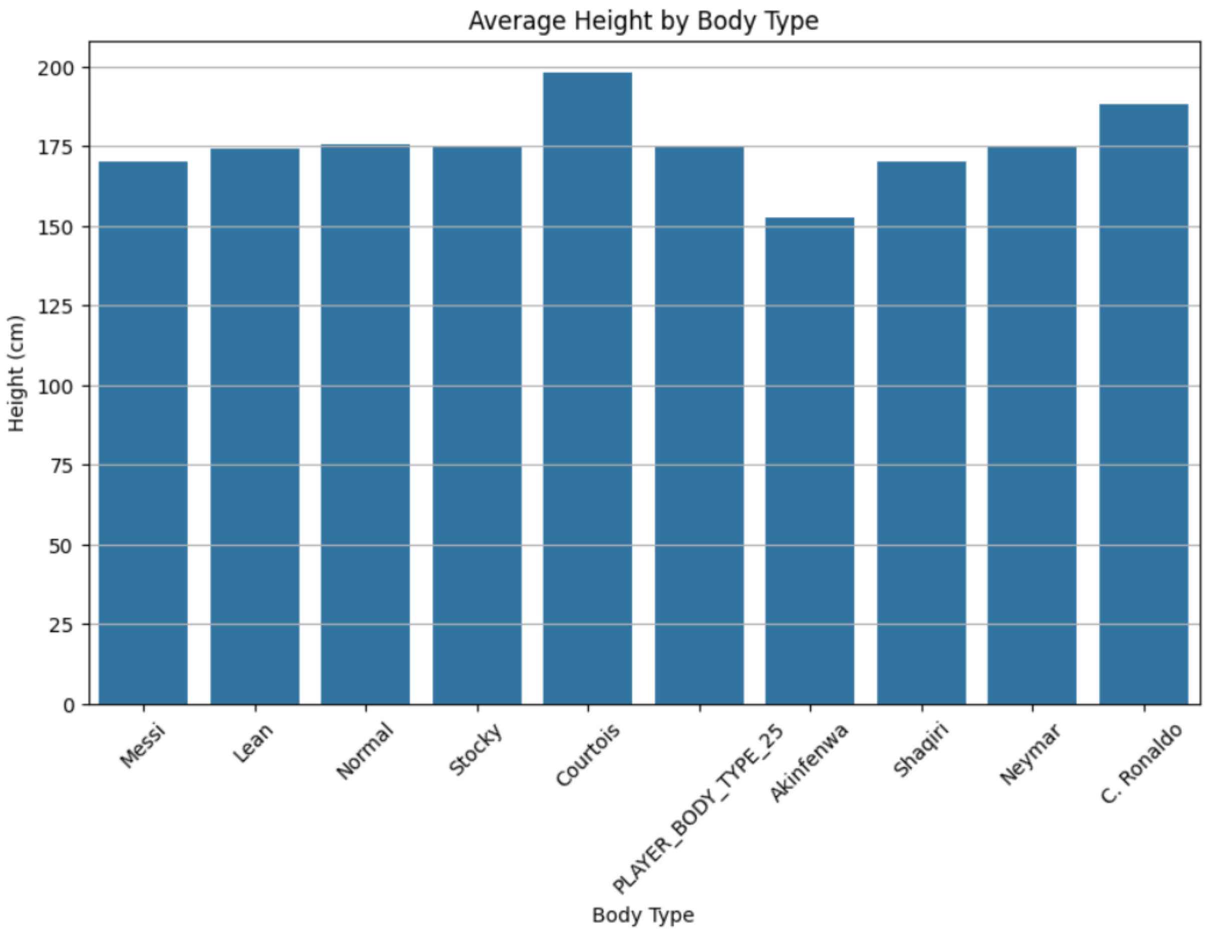
Both box plots use blue color for skill move ratings for both preferred foot cases. The choice of this color is neutral so it does not bring any form of bias or detract from what one should be primarily focused on: comparing distributions.

In conclusion, the design choices for both box plot and histogram were meticulously crafted to effectively communicate distribution of skill moves ratings and relationship between preferred foot and skill moves rating respectively. By leveraging basic perceptual tasks, using 2D space, arranging spatial layout and using an appropriate color scheme these visualizations facilitate simple comparisons, analysis of central tendencies, spread in data as well as understanding how our preferred feet relate closely with our ability to execute certain moves on football fields. Taken together, they offer a more comprehensive understanding of this dataset which leads to insightful analysis for decision-making contexts involving player attributes in Football.

3) BODY SHAPE AND PHYSICAL FEATURES:

Hypothesis: Specific physical attributes are best demonstrated by athletes of certain body shapes.

Bar graph:



Choice of Visualization:

It is suitable to use a bar chart when displaying and comparing average height across different categories like body types. This will accurately represent the quantitative measures of heights in this form of visualization.

Choice of Visual Design:

It is suitable to use a bar chart when displaying and comparing average height across different categories like body types. This will accurately represent the quantitative measures of heights in this form of visualization.

Body types are on x-axis while height (in cm) is on y-axis for two-dimensional space. Since every category is directly translated into its corresponding tallies, it promotes easier understanding and comparisons.

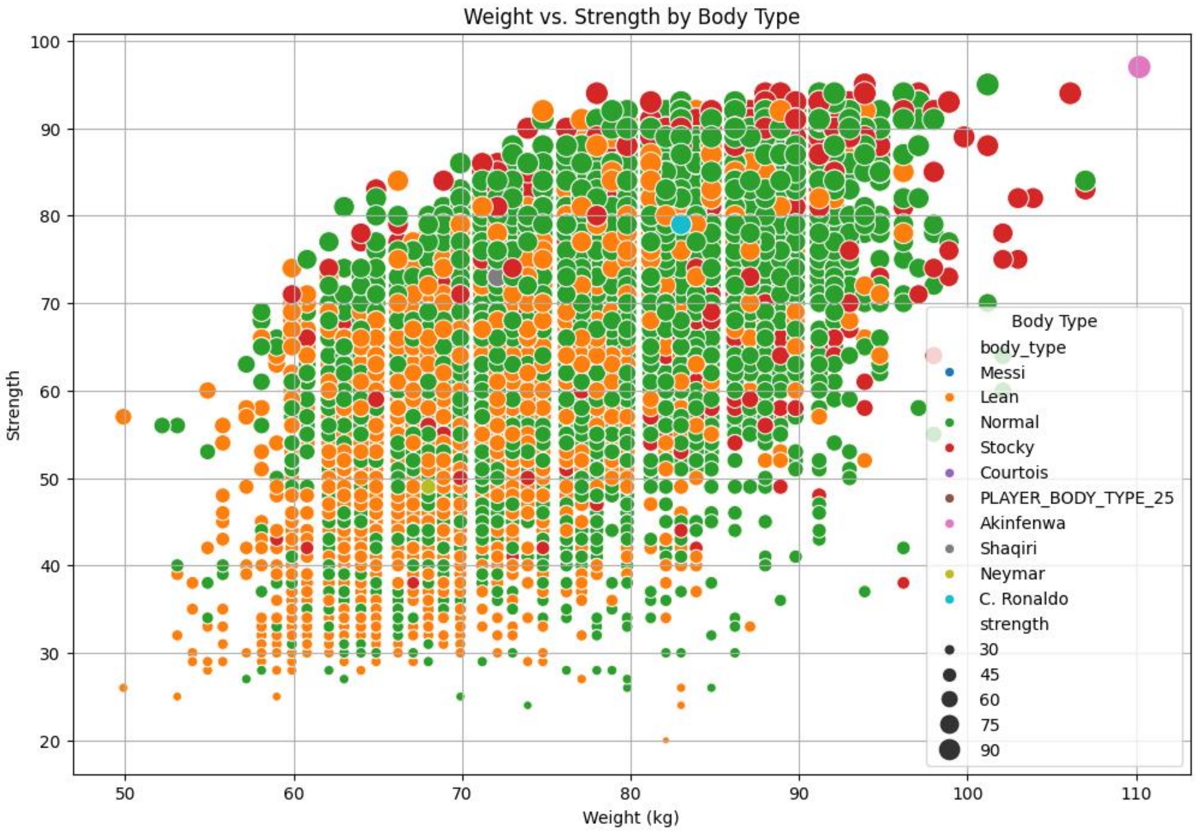
The bars appear horizontally next to one another, with each representing a different body type. Such positioning enables efficient visual comparison of all heights among all types within one plot area.

All bars are colored blue throughout, which makes them look neutral and egalitarian. Single colors aim at minimizing any possible bias or distraction from the main goal that entails comparing average heights across various body shapes.

Elementary Perceptual Task:

An example of an elementary perceptual task is provided by the bar chart used to compare lengths or heights. The height of each bar represents the average height value for each body type, thus allowing easy comparison of height differences between body types.

Bubble Graph:



Elementary Perceptual Task:

A Bubble plot represents an elementary perceptual task for comparing positions and patterns. Each point indicates a single player and their positions relative to one another as well as along weight –strength axes can be observed visually for any potential relationships or trends.

Choice of Visual Design:

Therefore, a bubble plot would be appropriate when illustrating relationships between two quantitative variables such as weight and strength within different categories such as body types. Bubble plots enable exploration for possible correlations or trends in data.

The two- dimensional space has weight on the x-axis and strength on the y-axis. This configuration allows representation of both variables simultaneously hence the compatibility of their relationship analysis.

Data points are scattered all over with their position determined by the respective weight and strength values. This arrangement shows how each player varies in terms of weight and strength for a given combination of the two variables.

Different colors can be used to represent different body types which would make it easier to discern and identify patterns or clusters within a group. The color choices differ widely from one another hence they allow for visual distinction of body types in bubble plot.

In summary, both charts effectively address the hypothesis by providing opportunities to examine possible correlations between physical attributes (height, weight, strength) with body types. The bar chart compares average heights across body types while scatter plot provides insights on the relationship between weight and strength among different body types.

Conclusion:

Overall, the FOOTBALL PLAYERS data analysis has provided important insights on how various attributes and features of a player are interconnected. This includes hypothesis-driven analysis as well as visualization techniques to see how age, preferred leg, somatotype and physical attributes shape players' performance or even future prospects

The most significant discoveries from the study include:

Age versus Potential: Younger players usually have higher potential ratings suggesting that age plays a major role in terms of what path a player will take in their career.

Preference for Foot and Skill Moves: There is reason to believe that playing with one's dominant foot can be associated with better average skill moves scores.

Body Type and Physical Attributes: Certain physical attributes such as height, weight and strength are associated with specific body types. For instance, when it comes to aerial duels taller players perform better while slimmer players are faster on the field.

To demonstrate these findings and explore underlying patterns within data, we effectively used different visualization techniques such as line charts, scatter plots, histograms, box plots, bar graphs and bubble graphs.

In summary this project has demonstrated how complex datasets can be analyzed using data visualization to derive insights meant for sports analytics decision-making processes.

Hypothesis-driven exploration and use of visualizations allows us understand more about factors contributing towards good performance by footballers.

Future work:

This work could be extended further through numerous additional studies so that more knowledge about players' characteristics can be gained. Some of them may include:

Additional Data Integration: Complementing FOOTBALL PLAYERS dataset with other sources like injury records of the player; match statistics or team performance metrics would help look at the whole picture around player's game.

Advanced Predictive Modeling: Developing predictive models based on machine learning methodologies would allow us to predict future potentials of different footballers according to historical figures. It could involve methods like regression analysis or decision trees or neural networks which identify leading indicators of success among players

Segmentation and Grouping of Players: Clustering algorithms can be used to group players according to their similarities in terms of characteristics or play style. Coaches and scouts may use these clusters for talent identification and player recruitment purposes

Player sentiment analysis: Analyzing social media sentiment as well as news articles about player performance might reveal some more aspects on how the public perceives them and how media talks about them; this could influence a player's worth in the market and his career path.

Interactive Visualization Tools for Data: Developing interactive dashboards or other visualizations that allow users interact with data more dynamically, thereby drilling down into particular performance metrics or player attributes would make it easier to use while allowing deeper analysis.

To further enhance our insights into player performance, it is necessary to extend the system by introducing these advanced techniques and more data sources in football analytics.

References:

1) Toemen, G. H. P. (2022). Player Performance Prediction in Football Using Machine Learning Techniques. https://pure.tue.nl/ws/portalfiles/portal/197522454/Thesis_BDS_Toemen

2) The influence of age on footballers' performance. (2022, December 13). Barça Innovation Hub. https://barcainnovationhub.fcbarcelona.com/blog/the-influence-of-age-on-footballers-performance/

3) Thomas, K. R. (2013). How a Quality Football Pitch Impacts the Quality, Skills and Technique of Footballers in Jamaica. *Methods*, 9. https://www.researchgate.net/profile/Kevin-Thomas-14/publication/335682261_How_does_a_football_pitch_impact_the_quality_skills_and_technique_of_footballers_in_Jamaica/links/5d74440992851cacdb293e26/How-does-a-football-pitch-impact-the-quality-skills-and-technique-of-footballers-in-Jamaica.pdf