FIFA - 19 Player Statistics

ANOOP JOHN ALOCIOUS
ID : 31084354

Introduction -

Football is considered to be the world's most popular sport, as it is followed by millions of fans around the world. Most of the top football nations arise from Europe, with the competition level significantly higher in Europe as compared to all the other continents. Each country has their own league with around 20+ teams competing to win the league. Each team in a league could have players of other nationalities.

FIFA 19 is a simulated football video game developed by Electronic Arts Sports (EA Sports). EA Sports is a subdivision of Electronic Arts which produces sports video games. The game mainly consists of teams , and players that have played in different leagues until 2018. There are various modes in FIFA - 19, such as Classic mode, Manager mode. Classic mode is where the users can pick 2 teams and play against each other. Manager mode allows users to act as the manager of a preferred team and manage their team according to their interest.

Motivation -

I have been a huge admirer of Football since my childhood. Initially, I used to watch live matches on television, and later I started to go and play football outdoors. With EA Sports releasing different versions of FIFA every year, it got me excited to know more about the statistics of each team and their players. I was curious to explore how each attribute of a player or team influenced other aspects of the game such as Ratings, Potential etc. Hence, I decided to perform a deep analysis on each of these factors.

Problem Description -

I wish to perform analysis and answer few of the questions provided below:

1. Is there any correlation between the age of a player on his overall rating and value? Also, Does height and weight affect the overall rating as well?

2. Which nations produce the most number of players? Is there any relationship between the wage of a player and their nationality?

3. Which team has the highest potential? Is there any relationship between the potential of a player and his wages?

I have performed my analysis using tools like RStudio, Tableau, which have helped me answer the proposed questions.

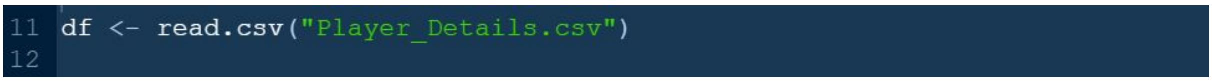
Data Description -

Dataset can be found in:
https://drive.google.com/file/d/1iP2eYRLKd0x7IHsSr0D8EBu5N0s0j_tZ/view?usp=sharing

The dataset has 18206 rows & 89 columns. Each row consists of characters, numbers and URLs. Each column represents the attribute of a player such as Overall Rating, Potential, etc. The dataset also contains positional information of each player as well as their wages and value.

Data Reading -

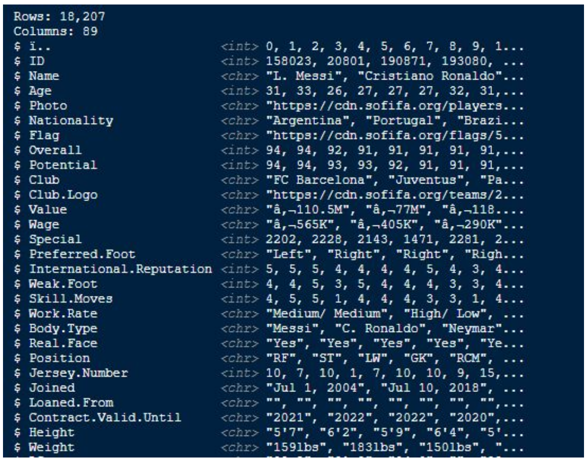
As we know, R is a powerful analytical tool which can be effectively used to check for anomalies for a given dataset. So, I have used RStudio to read the data and analyse it before further exploration.



```
df <- read.csv("Player_Details.csv")
```

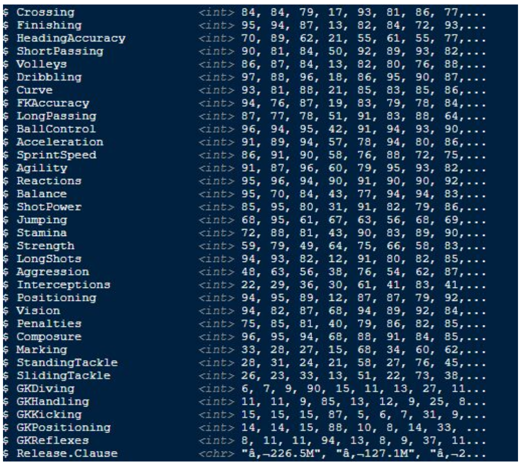
Figure 1: Reading Data

Once the data is read into RStudio, a glimpse of information regarding the names and type of each column is done.



```
Rows: 18,207
Columns: 89
$ ï..                      <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 1...
$ ID                       <int> 158023, 20801, 190871, 193080, ...
$ Name                     <chr> "L. Messi", "Cristiano Ronaldo"...
$ Age                      <int> 31, 33, 26, 27, 27, 27, 32, 31,...
$ Photo                    <chr> "https://cdn.sofifa.org/players...
$ Nationality              <chr> "Argentina", "Portugal", "Brazi...
$ Flag                     <chr> "https://cdn.sofifa.org/flags/5...
$ Overall                  <int> 94, 94, 92, 91, 91, 91, 91, 91,...
$ Potential                <int> 94, 94, 93, 93, 92, 91, 91, 91,...
$ Club                     <chr> "FC Barcelona", "Juventus", "Pa...
$ Club.Logo                <chr> "https://cdn.sofifa.org/teams/2...
$ Value                    <chr> "â,¬110.5M", "â,¬77M", "â,¬118....
$ Wage                     <chr> "â,¬565K", "â,¬405K", "â,¬290K"...
$ Special                  <int> 2202, 2228, 2143, 1471, 2281, 2...
$ Preferred.Foot           <chr> "Left", "Right", "Right", "Righ...
$ International.Reputation <int> 5, 5, 5, 4, 4, 4, 4, 5, 4, 3, 4...
$ Weak.Foot                <int> 4, 4, 5, 3, 5, 4, 4, 4, 3, 3, 4...
$ Skill.Moves              <int> 4, 5, 5, 1, 4, 4, 4, 3, 3, 1, 4...
$ Work.Rate                <chr> "Medium/ Medium", "High/ Low", ...
$ Body.Type                <chr> "Messi", "C. Ronaldo", "Neymar"...
$ Real.Face                <chr> "Yes", "Yes", "Yes", "Yes", "Ye...
$ Position                 <chr> "RF", "ST", "LW", "GK", "RCM", ...
$ Jersey.Number            <int> 10, 7, 10, 1, 7, 10, 10, 9, 15,...
$ Joined                   <chr> "Jul 1, 2004", "Jul 10, 2018", ...
$ Loaned.From              <chr> "", "", "", "", "", "", "", "",...
$ Contract.Valid.Until     <chr> "2021", "2022", "2022", "2020",...
$ Height                   <chr> "5'7", "6'2", "5'9", "6'4", "5'...
$ Weight                   <chr> "159lbs", "183lbs", "150lbs", "...
```

Figure 2: Data Glimpse(First few columns)

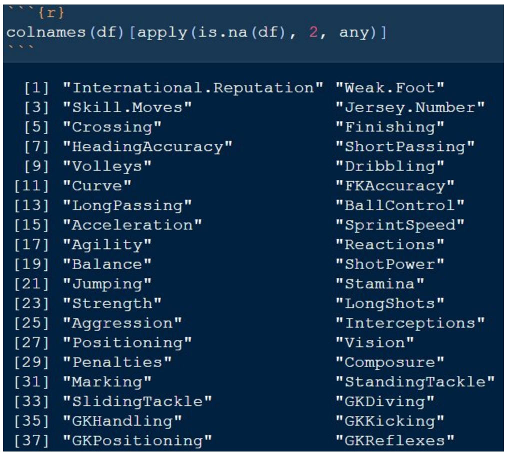


```
$ Crossing                 <int> 84, 84, 79, 17, 93, 81, 86, 77,...
$ Finishing                <int> 95, 94, 87, 13, 82, 84, 72, 93,...
$ HeadingAccuracy          <int> 70, 89, 62, 21, 55, 61, 55, 77,...
$ ShortPassing             <int> 90, 81, 84, 50, 92, 89, 93, 82,...
$ Volleys                  <int> 86, 87, 84, 13, 82, 80, 76, 88,...
$ Dribbling                <int> 97, 88, 96, 18, 86, 95, 90, 87,...
$ Curve                    <int> 93, 81, 88, 21, 85, 83, 85, 86,...
$ FKAccuracy               <int> 94, 76, 87, 19, 83, 79, 78, 84,...
$ LongPassing              <int> 87, 77, 78, 51, 91, 83, 88, 64,...
$ BallControl              <int> 96, 94, 95, 42, 91, 94, 93, 90,...
$ Acceleration             <int> 91, 89, 94, 57, 78, 94, 80, 86,...
$ SprintSpeed              <int> 86, 91, 90, 58, 76, 88, 72, 75,...
$ Agility                  <int> 91, 87, 96, 60, 79, 95, 93, 82,...
$ Reactions                <int> 95, 96, 94, 90, 91, 90, 90, 92,...
$ Balance                  <int> 95, 70, 84, 43, 77, 94, 94, 83,...
$ ShotPower                <int> 85, 95, 80, 31, 91, 82, 79, 86,...
$ Jumping                  <int> 68, 95, 61, 67, 63, 56, 68, 69,...
$ Stamina                  <int> 72, 88, 81, 43, 90, 83, 89, 90,...
$ Strength                 <int> 59, 79, 49, 64, 75, 66, 58, 83,...
$ LongShots                <int> 94, 93, 82, 12, 91, 80, 82, 85,...
$ Aggression               <int> 48, 63, 56, 38, 76, 54, 62, 87,...
$ Interceptions            <int> 22, 29, 36, 30, 61, 41, 83, 41,...
$ Positioning              <int> 94, 95, 89, 12, 87, 87, 79, 92,...
$ Vision                   <int> 94, 82, 87, 68, 94, 89, 92, 84,...
$ Penalties                <int> 75, 85, 81, 40, 79, 86, 82, 85,...
$ Composure                <int> 96, 95, 94, 68, 88, 91, 84, 85,...
$ Marking                  <int> 33, 28, 27, 15, 68, 34, 60, 62,...
$ StandingTackle           <int> 28, 31, 24, 21, 58, 27, 76, 45,...
$ SlidingTackle            <int> 26, 23, 33, 13, 51, 22, 73, 38,...
$ GKDiving                 <int> 6, 7, 9, 90, 15, 11, 13, 27, 11...
$ GKHandling               <int> 11, 11, 9, 85, 13, 12, 9, 25, 8...
$ GKKicking                <int> 15, 15, 15, 87, 5, 6, 7, 31, 9,...
$ GKPositioning            <int> 14, 14, 15, 88, 10, 8, 14, 33, ...
$ GKReflexes               <int> 8, 11, 11, 94, 13, 8, 9, 37, 11...
$ Release.Clause           <chr> "â,¬226.5M", "â,¬127.1M", "â,¬2...
```

Figure 3: Data Glimpse(Last few columns)

Missing Values -

Since there are thousands of rows, it is practically impossible to manually check for missing data. Using R, we can find and filter the missing values as it could affect the accuracy of our results.



~~~
```{r}
colnames(df)[apply(is.na(df), 2, any)]
```

 [1] "International.Reputation" "Weak.Foot"
 [3] "Skill.Moves"              "Jersey.Number"
 [5] "Crossing"                 "Finishing"
 [7] "HeadingAccuracy"          "ShortPassing"
 [9] "Volleys"                  "Dribbling"
[11] "Curve"                    "FKAccuracy"
[13] "LongPassing"              "BallControl"
[15] "Acceleration"             "SprintSpeed"
[17] "Agility"                  "Reactions"
[19] "Balance"                  "ShotPower"
[21] "Jumping"                  "Stamina"
[23] "Strength"                 "LongShots"
[25] "Aggression"               "Interceptions"
[27] "Positioning"              "Vision"
[29] "Penalties"                "Composure"
[31] "Marking"                  "StandingTackle"
[33] "SlidingTackle"            "GKDiving"
[35] "GKHandling"               "GKKicking"
[37] "GKPositioning"            "GKReflexes"
~~~

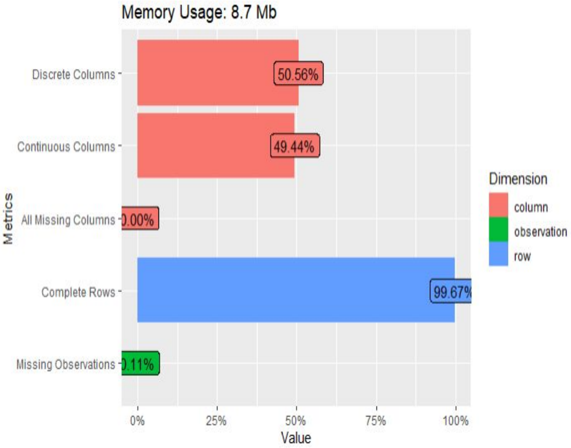
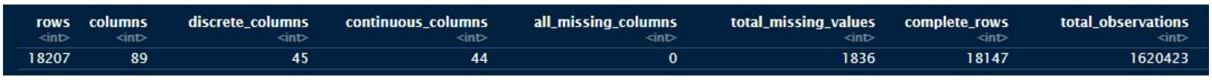


Figure 4: Columns having missing values *Figure 5: Missingness of the Data*

It can be observed that around 0.11% of the observations consist of missing values. When we are considering this massive dataset, even a minute percentage of missing values might have a huge impact on our results. It can be noted that 38 columns in our dataset consists of missing values. However, for our analysis these columns are irrelevant, so we have dropped them.

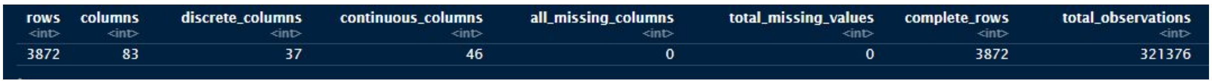
- **Before filtering the observations with missing values:**



rows <int>	columns <int>	discrete_columns <int>	continuous_columns <int>	all_missing_columns <int>	total_missing_values <int>	complete_rows <int>	total_observations <int>
18207	89	45	44	0	1836	18147	1620423

We can see that there are 1836 missing values in our dataset, and there is no column that consists of full empty values.

- **After filtering the observations with missing values:**

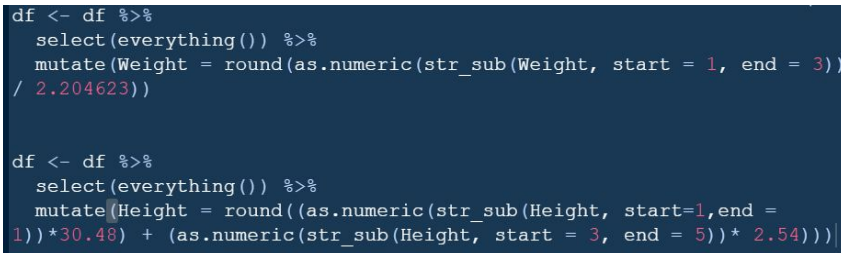


rows <int>	columns <int>	discrete_columns <int>	continuous_columns <int>	all_missing_columns <int>	total_missing_values <int>	complete_rows <int>	total_observations <int>
3872	83	37	46	0	0	3872	321376

After dropping the irrelevant columns which are not needed for our analysis, we have 3872 valid observations without any missing values.

Data Manipulation -

Some columns important for our analysis are converted to relevant formats for the ease of calculation. It can be noted that **"Value"** and **"Wage"** are given in character format preceded with a currency symbol. We have altered this column by removing the currency symbol and have also converted them into their respective values in thousands. Height and Weight fields were also modified as they were provided in feet and lbs respectively. These values are converted to centimeters and kilograms as they are used as the standard units of measurements for our analysis.



```
df <- df %>%
  select(everything()) %>%
  mutate(Weight = round(as.numeric(str_sub(Weight, start = 1, end = 3))
/ 2.204623))


df <- df %>%
  select(everything()) %>%
  mutate(Height = round((as.numeric(str_sub(Height, start=1,end =
1))*30.48) + (as.numeric(str_sub(Height, start = 3, end = 5))* 2.54)))
```

Figure 6: Conversion of Height and Weight

Exploration -

- **Relationship between Age and Overall rating of a Player**

Value of a player is considered as the price of the player based on the contract signed by the respective player for a certain club. We are going to analyse if the age of a player has any impact on his value.

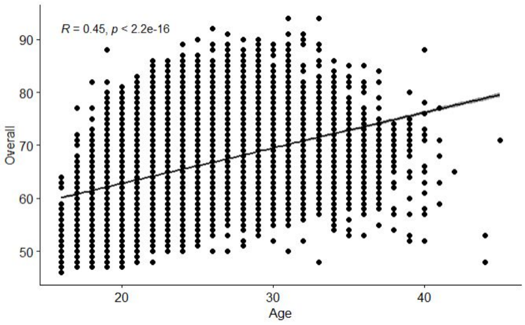


Figure 7: Overall Rating vs Age

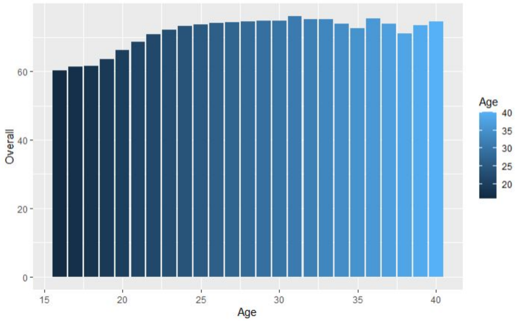


Figure 8: Bar graph

It is observed that age and overall rating of a player are positively correlated until 30 years of age. For all the players above 30, this trend is not witnessed as the rating might depend on other factors such as performance etc.

We have used Pearson's correlation coefficient to verify our analysis. We have computed 0.45 as the R-Value(correlation coefficient), which means there is some correlation between Age and Overall rating of a player. A positive correlation implies that as one variable increases the other one tends to increase as well.

- **Relationship between Age and Value of a Player**

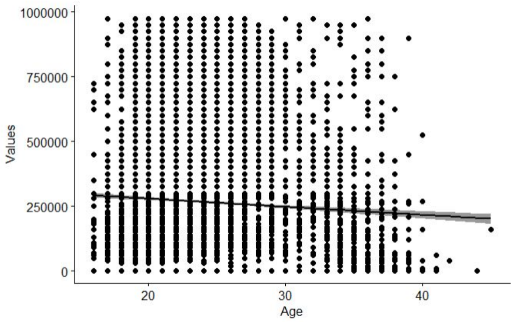


Figure 9: Age vs Value

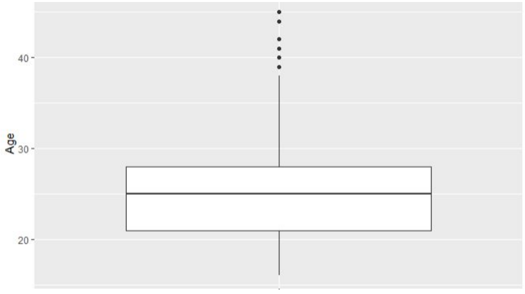


Figure 10: Boxplot for Age

Using the boxplot, it can be noted that most of the players were under 30 years of Age. We can also see that there are some outliers, few players having age between 38-45. By plotting the graph with all the observations, and by drawing a regression line, it can be observed that the Value of a player slightly tends to decrease as the age of a player increases. This decay in value is more when the player is over 30 years old.

- **Relationship between Height, Weight and Overall Rating**

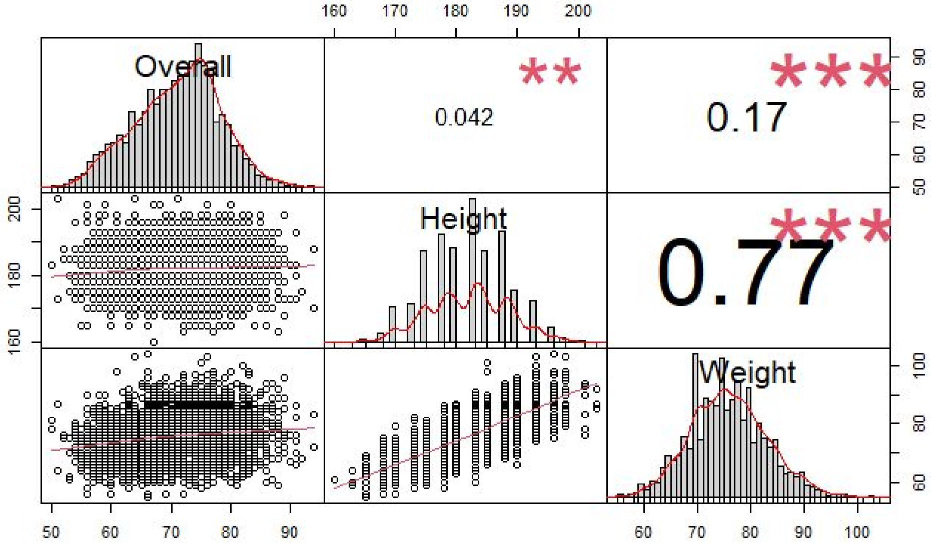


Figure 11: Correlation Matrix for Height, Weight & Overall Rating

To identify the relationship between variables, we can use the correlation matrix. Correlation matrix provides us the correlation value between each variable inside a matrix. The stars represent the significance levels, which is used to show how likely a pattern between the given variables is meant to happen.

By analyzing the correlation matrix, we can find that the correlation between Height and Overall rating is close to zero. Hence, it is safe to conclude that Height and Overall Rating are not correlated. Whereas, from the bar graphs we can see that the Overall Rating increases with the increase in Weight of a player.

- **Top 10 Footballing Nations**

There are players from over 100 countries who play for different clubs. We are interested to see the top 10 countries that produce football talents around the world.

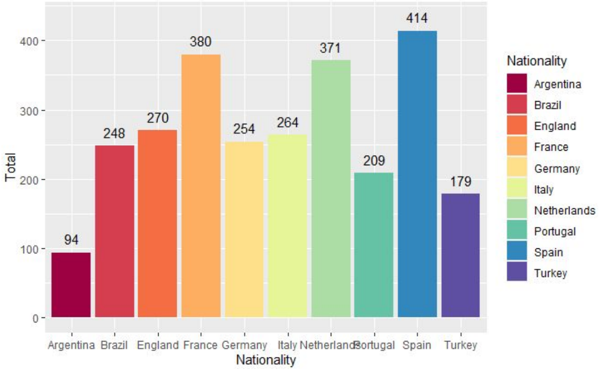


Figure 12: Bar Graph for Top 10 Countries

We can see that nearly 2000 players arise from the top 10 football nations in the world. Majority of the players belong to Spain, followed by France and Netherlands. Another eye-catching information that we can perceive from the figure is that 8 out of the 10 countries are European countries. So, we can also conclude that competition is higher in Europe as compared to other continents.

- **Relationship between Wage of a player and Nationality**

Each player is entitled to a certain wage which is paid to them by the club on a weekly basis. There has been a notion that football players from the top footballing nations tend to be paid more than players arising from other nations. We would like to analyse if this fact is true and see if the wages of a player depends on his nationality.

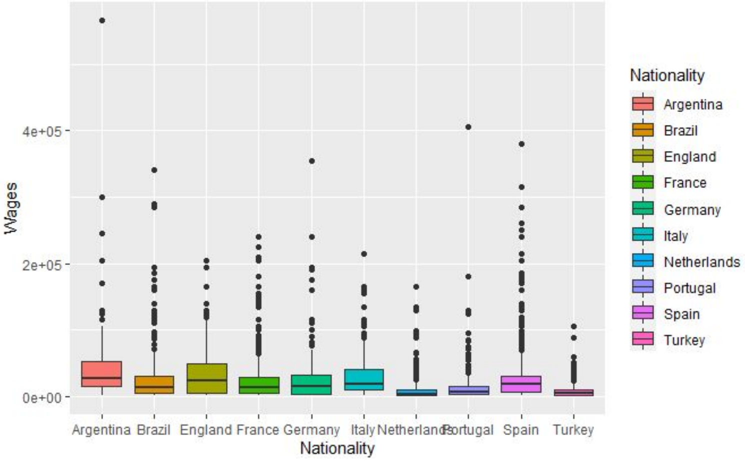


Figure 13: Boxplot for Nationality & Wages

From the above figure, we can see that for each player arising from the top 10 nations does not really intend to have higher wages. Considering Spain, they have emerged as the nation which produces the most number of players. The median wages value of Spain is comparatively less than that of England and Argentina. So, it is clear that the wages of a player is independent of his nationality, but it may depend on the player's individual attributes such as Overall Rating etc.

- **Team with the highest Potential**

There are 2 types of teams, one is the national team where players from the same nationality combine and play as a team, and the other is the clubs, where players from different nations come together and compete in a certain league. Each player has a potential, which is nothing but how much a player can grow given his current form or abilities. We are going to perform analysis in order to find out which all clubs and national teams have high potential players.

For this analysis, we have used Tableau to visualize the result. Tableau is an open-source software which can be used easily to create visualization. We have visualized the result using a tree-graph, which lists each nationality having the highest potential to the lowest potential.

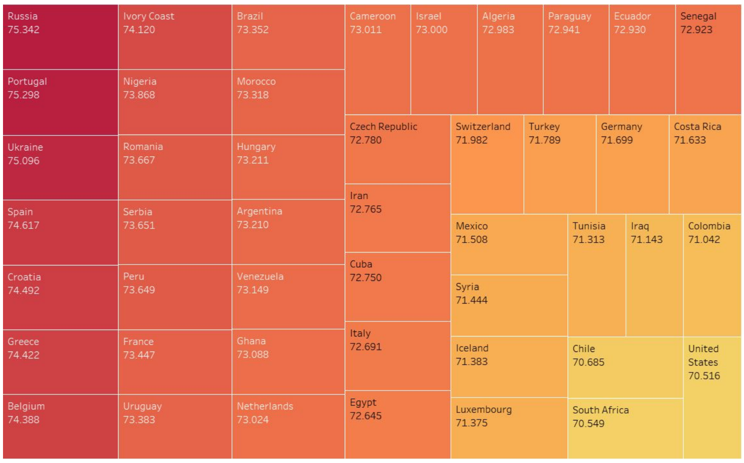


Figure 14: Footballing Nations and Average Potential

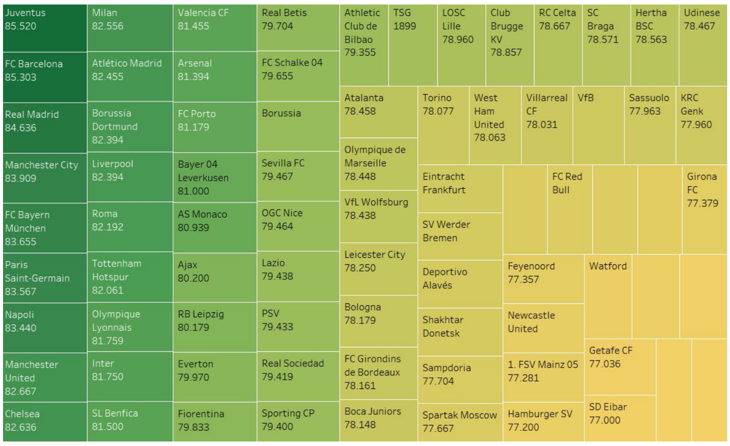


Figure 15: Football Clubs and Average Potential

From our visualization, it is clear that countries such as Spain, France, that produces the majority of the football players do not necessarily have high potential. Countries like Russia, Portugal tend to have players with more potential than these. Clubs like Juventus, Barcelona, Real Madrid, Manchester City have players with high potential.

- **Relationship between Wages and Potential**

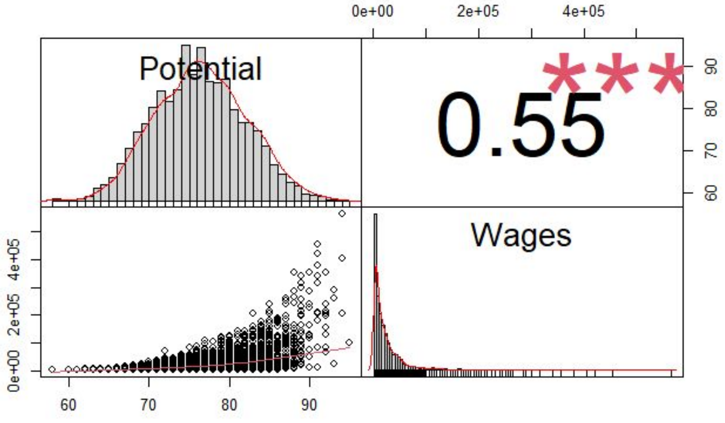


Figure 16: Correlation Matrix of Potential and Wages

From the correlation matrix, we can observe that the wages of a player has a correlation of 0.55 with the player's potential. In other words, as the potential of the player increases, the wages of the player also increases.

Conclusion -

From the above exploration section, we can conclude that:

- Age affects the overall rating and value of a player
- Height does not have an impact on the overall rating of a player
- Weight of a play does impact the overall rating of a player
- Top football nations are Spain, France, Netherlands
- Majority of the top football nations are European countries
- Russia, Portugal are nations that have players with high potential
- Juventus, Barcelona, Real Madrid are clubs that have players with high potential
- Wages of a player does not depend on the nationality of the player
- Potential of a player and his wages are directly proportional to each other

Reflection -

We have reached our conclusion by using different analytical tools on the dataset. We were able to see the correlation between different variables on the dataset. After performing all the analysis, we can understand how some attributes of a player could impact their value and wages. We were able to clear the wrong notion about the relationship between the wage of a player and his nationality using visualization. Hence, exploration and visualization of the data can be considered as the most integral part of data analytics.

References -

- Software Used -
    1. Tableau Desktop Public Edition (Version 2020.3)[Windows]. Christian Chabot, Pat Hanrahan and Chris Stolte, in Mountain View, California. Downloaded from https://www.tableau.com/
    2. RStudio Version 1.3.1056 © 2009-2020 RStudio, PBC
"Water Lily" (5a4dee98, 2020-07-07) for Windows

- R Packages -

    1. tidyverse: Easily Install and Load the 'Tidyverse'. Version: 1.3
https://cran.r-project.org/web/packages/tidyverse/index.html

    2. dplyr: A Grammar of Data Manipulation. Version: 1.0.2
https://cran.r-project.org/web/packages/dplyr/index.html

    3. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. Version: 3.3.2
https://cran.r-project.org/web/packages/ggplot2/index.html