MINI PROJECT REPORT - EDA + STATS

EDA AND STATISTICAL ANALYSIS OF FIFA CASE STUDY

GROUP 3

CONTENTS

Topics

1. Team Members
2. Introduction to the Problem Statement
3. Technologies Used
4. Skills Used and Developed
5. Data Description
6. Data Collection and Cleaning
7. Problem Solving Steps
8. Takeaways and Conclusions
9. Future Steps

TEAM MEMBERS

Aditya Aryan

Bhavani Kanaparthi

Naman Goswami

Neha Mondal

INTRODUCTION TO THE PROBLEM STATEMENT

The full form for FIFA is Fédération Internationale de Football Association. It is the most popular international organisation for the widely known sport of football. FIFA is a grand body composed of great many member federations which manages numerous football competitions and tournaments all around the world, with the FIFA World Cup being the most awaited and respected among them. It is also well-known for the creation of the rules and laws related to the sport of football like the proportions of a football field or the count of players in a football team. The enrolment of football players and coaches also comes under the vast list of functions of FIFA. The association constantly yearns to make efforts to develop the sport and ensure that it is played fair and safe in the years to come.

For this case study, we are told of the introduction of a new football club named ‘Brussels United FC’. It does not yet have a team. As the Data Science Team, we are given the task to use a data-based approach for the construction of a report which would help recommend players for the main team of the club. Fifteen players are required for the formation of the team. The budget for hiring of the players is fixed. The management wants to choose from a total of twenty possible players with potential. For the job, we are provided with a large data of players acquired through FIFA.

TECHNOLOGIES USED

While the theoretical aspects and knowledge and experience gained from a project are significant, no Data Science project can be finished without the use of technology. It is essential to fulfil the asks of the problem statement. Hence, it is important we address the non-theoretical sides behind the outcomes of this project.

We have primarily come across two types of technologies throughout the creation of our project. One of them is the programming language we have used for various calculations, data cleaning and visualization etc. while the other is the platform which provided us with the facility of making use of the said programming language. They are listed as follows:

✓ **Python (Programming Language)**

Python is a high-level, interpreted and general-purpose programming language. It is often known as the "Swiss Army Knife" of programming languages. Commended for its simplicity, readability and versatility, it uses a very simple and easy to understand syntax which makes it a perfect choice for beginners to programming.

✓ **Jupiter Notebook (Integrated Development Environment)**

Jupiter Notebook is a web-based programming environment which is open-source. It provides its users the ability to create and edit programs, graphs and comments all in a single document. It is a very popular tool used by data scientists and mentors of data analysis and machine learning. It is used to code in various programming languages, including Python, R, Julia, and many others.

SKILLS USED AND DEVELOPED

With the theoretical knowledge, one must also possess the skills which are important for the application of the necessary steps during a project. While most of the basic skills are applied, others are also developed by learning new techniques and taking different approaches to a problem.

The skills we have used are as follows:

- ✓ **Programming:** Basic Python programming has helped us in using libraries, doing calculations by using formulas and looping etc. throughout the timeline of the project.
- ✓ **Inferencing:** In the conceptual as well as the dataset-based part, we have made numerous inferences by looking at numbers and graphs.
- ✓ **Data Pre-processing:** We have performed outlier detection, dropping and conversion of columns, null removal and imputation on the data provided.
- ✓ **Data Visualization:** Various graphs and plots have been created from our knowledge of visualization to dig deeper into the data for information.
- ✓ **Statistical Analysis:** We have used statistical concepts for finding out probabilities and testing hypotheses and claims stated in the questions.
- ✓ **Domain Knowledge:** For the dataset-based part, the domain knowledge has been used to make significant decisions at times.
- ✓ **Exploratory Data Analysis:** To analyse the data for hidden information and to identify patterns and insights from the data, we have made use of the concepts of EDA.
- ✓ **Problem-Solving:** While we were mostly told exactly what to do to achieve results, it took us to solve undefined problems at instances.

Lastly, the application of above-mentioned skills has helped us in their development on some level.

DATA DESCRIPTION

The data set we are provided with is filled with information about different players, the clubs they are playing for, and a vast list of columns which can be used to find out players with good performance.

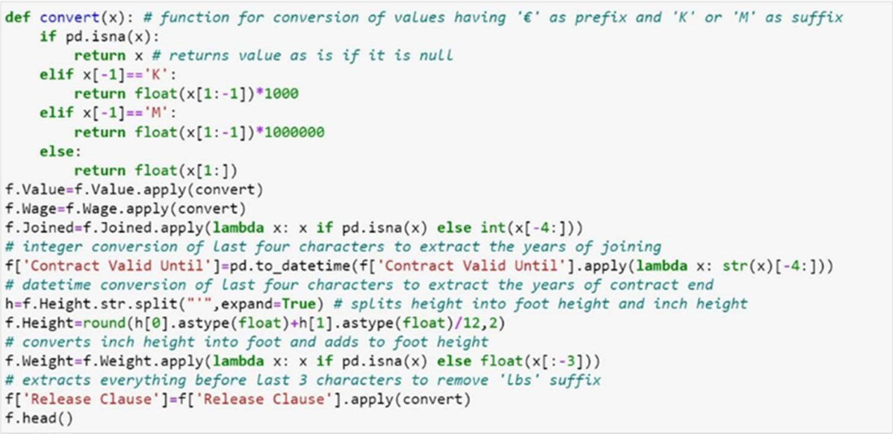
There are 25,490 records and a total of 60 columns in the data. 18 of the columns are originally categorical while the remaining 42 are numerical. 11,833 records were found to be duplicated, and several columns have missing values going up to a maximum count of 1,171. The data dictionary is as follows:

Column	Description	Column	Description
ID	unique ID for every player	Dribbling	rating on scale of 100
Name	name	Curve	rating on scale of 100
Age	age	FKAccuracy	rating on scale of 100
Photo	URL to the players' photo	LongPassing	rating on scale of 100
Nationality	nationality	BallControl	rating on scale of 100
Flag	URL to players' country flag	Acceleration	rating on scale of 100
Overall	overall rating	SprintSpeed	rating on scale of 100
Potential	potential rating	Agility	rating on scale of 100
Club	current club	Reactions	rating on scale of 100
Club Logo	URL to club logo	Balance	rating on scale of 100
Value	current market value	ShotPower	rating on scale of 100
Wage	current wage	Jumping	rating on scale of 100
Preferred Foot	left/right	Stamina	rating on scale of 100
International Reputation	rating on scale of 5	Strength	rating on scale of 100
Weak Foot	rating on scale of 5	LongShots	rating on scale of 100
Skill Moves	rating on scale of 5	Aggression	rating on scale of 100
Work Rate	attack work rate / defence work rate	Interceptions	rating on scale of 100
Body Type	body type of player	Positioning	rating on scale of 100
Position	position on the pitch	Vision	rating on scale of 100
Jersey Number	jersey number	Penalties	rating on scale of 100
Joined	joined date	Composure	rating on scale of 100
Loaned From	club name if applicable	Marking	rating on scale of 100
Contract Valid Until	contract end date	StandingTackle	rating on scale of 100
Height	height of the player	SlidingTackle	rating on scale of 100
Weight	weight of the player	GKDiving	rating on scale of 100
Crossing	rating on scale of 100	GKHandling	rating on scale of 100
Finishing	rating on scale of 100	GKKicking	rating on scale of 100
HeadingAccuracy	rating on scale of 100	GKPositioning	rating on scale of 100
ShortPassing	rating on scale of 100	GKReflexes	rating on scale of 100
Volleys	rating on scale of 100	Release Clause	release clause value

DATA COLLECTION AND CLEANING

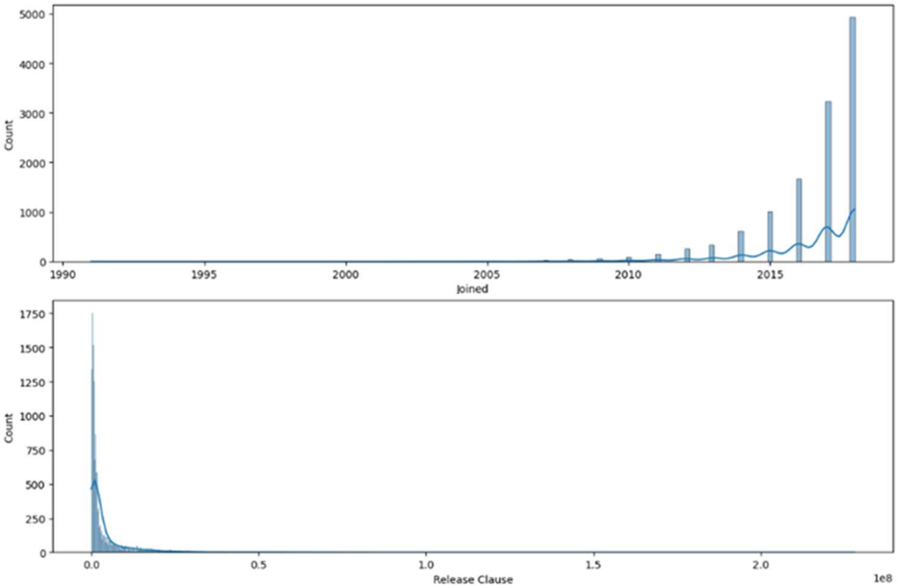
The data was collected from the “fifa.csv” file we were provided with.

For cleaning the data, we performed various steps starting with the removal of redundant columns like 'Photo', 'Flag', 'Club', 'Club Logo', 'Jersey Number', and 'Loaned From'. Various variables needed to be converted to extract information from them like 'Value', 'Wage', 'Joined', 'Contract Valid Until', 'Height', 'Weight' and 'Release Clause'. We used different techniques for their conversion.



```
def convert(x): # function for conversion of values having '€' as prefix and 'K' or 'M' as suffix
    if pd.isna(x):
        return x # returns value as is if it is null
    elif x[-1]=='K':
        return float(x[1:-1])*1000
    elif x[-1]=='M':
        return float(x[1:-1])*1000000
    else:
        return float(x[1:])
f.Value=f.Value.apply(convert)
f.Wage=f.Wage.apply(convert)
f.Joined=f.Joined.apply(lambda x: x if pd.isna(x) else int(x[-4:]))
# integer conversion of last four characters to extract the years of joining
f['Contract Valid Until']=pd.to_datetime(f['Contract Valid Until'].apply(lambda x: str(x)[-4:]))
# datetime conversion of last four characters to extract the years of contract end
h=f.Height.str.split("'",expand=True) # splits height into foot height and inch height
f.Height=round(h[0].astype(float)+h[1].astype(float)/12,2)
# converts inch height into foot and adds to foot height
f.Weight=f.Weight.apply(lambda x: x if pd.isna(x) else float(x[:-3]))
# extracts everything before last 3 characters to remove 'lbs' suffix
f['Release Clause']=f['Release Clause'].apply(convert)
f.head()
```

The duplicated records were dropped from the data. We removed the records for columns with smaller amounts of missing values i.e., less than 2% while imputed for the rest with mode in categorical and median in numerical columns after checking their distributions.



The ‘Body Type’ column was found to have some invalid values on inspection, which were dropped.

PROBLEM SOLVING STEPS

Part – A

1. We converted the list of BMIs into a pandas Series and used mean(), mode() and median() functions.
2. Used max()-min() for range and var() and std() for variance and standard deviation respectively.
3. Used the formula for mean absolute deviation.
4. Used the formula for Pearson's coefficient of skewness.
5. We found values greater than or equal to mean minus standard deviation and less than or equal to mean plus standard deviation and calculated the length of the resulting Series.
6. We have calculated the quartiles using quantile().
7. We checked for values which are more than upper limit and less than lower limit for checking outliers.
8. Used plot() to plot the box plot.
9. Made a Data Frame with BMI and ranks giving 'pct=True' and then got rank for BMI = 25.
10. Calculated probability by dividing the number of adults with BMI > 25 by the total number of adults.
11. Used seaborn histplot() and given parameter 'kde = True' which adds Kernel Density Estimate (KDE) to the histogram.
12. Used seaborn histplot() with 'stat = probability' as a parameter to get probability distribution.
13. We used Kernel Density Estimate (KDE) plot using seaborn library to check distribution of data. We have taken 100 samples of sample size 5 with replacement, calculated their means and stored into list, and then plotted the KDE for the list. Repeated the steps inside another loop to iterate over sample size from 10 to 30.
14. Used probability mass function to find probability for exactly 6 people with probability calculated in question 10.
15. As both np and nq were more than 5, we considered normal approximation. Calculated mean using formula of mean of a binomial distribution and standard deviation using square root of formula of variance of a binomial distribution. Cumulative distribution functions for normal distribution use interval of 49.5 to 50.5 considering continuity correction of 0.5 on 50.
16. Used interval() from stats.norm to get the interval and used z-distribution as n > 30.
17. Used formula for sample size and then rounded to next nearest integer.
18. Used the same formula with different values.
19. Performed One Proportion Z-test and returned p-value.
20. Performed Two Proportion Z-test and returned p-value calculating number of successes from percentages.

Part – B

1. Imported the necessary libraries and read the csv file.
2. Dropped the unnecessary columns by using drop() with parameter ‘axis = 1’.
3. For ‘Wage’, 'Release Clause’ and ‘Value’, we defined a function to convert ‘K’ and 'M’ into 1000 and 1000000 and kept null unchanged. For ‘Joined’ and ‘Contract Valid Until’, we took last 4 characters and converted them into int using apply() function. Same treatment done with ‘Weight’ column here to take all characters excluding the last 3, for ‘Height’ column we used split() to get foot and inches separately and then converted them to foot.
4. Used duplicated() to check duplicates and drop_duplicates() to drop the duplicates.
5. Used a for loop to iterate over the numerical columns of the data set and then found variance and STD of each.
6. Found count and percentage of missing values for all the columns. Dropped the rows with missing values in ‘Contract Valid Until’. Skewness and kurtosis of ‘Joined’ and ‘Release Clause’ were checked using skew() and kurt() respectively. Plotted the box and KDE plot for both the columns. We imputed missing values of ‘Joined’ with its mode and the missing values of ‘Release Clause’ with its median.
7. First, we calculated the Q1, Q3 and IQR for all numerical columns. Then using them, we checked in the data for unique names for which all numerical data is between 2.0 * IQR – Q1 and 2.0 * IQR + Q3.
8. We checked value counts then dropped garbage values for ‘Body type’. Then plotted count plot for categorical columns to check for cardinality.
9. Used pairplot().
10. Made a Data Frame with ‘Contract Valid Until’ = 2020-01-01 and used nlargest() with parameter ‘number = 20’ and “columns = ‘Overall’”. Used regplot() to plot correlation plot.
11. Took a for loop to iterate over all unique positions and then used the method above to get the top 5 player names. Made a Series and then used a for loop which iterates over different unique values of position column to calculate the mean wage and updated in the Series.

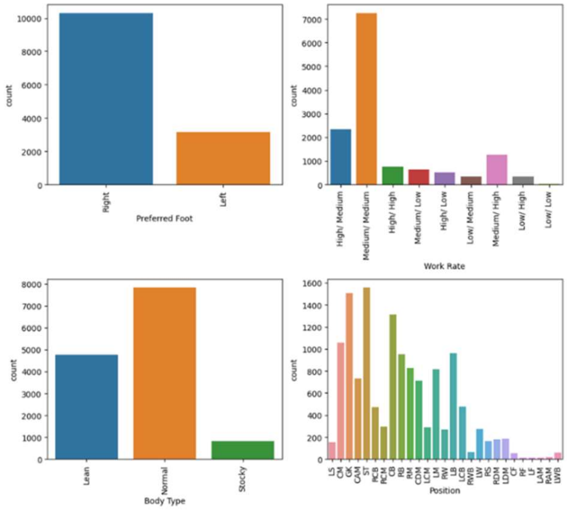
Statistical Analysis

1. We separated right-footed and left-footed players. Then we plotted histograms for the ‘Overall’ column for both sets and checked the skewness. We used Two Sample Independent T-test with the two sets we made before and got the p-value.
2. We performed the same test in this question as well. The data we provided here is one set with potential of age greater than 35 and other less than 35 and got the p-value.
3. We used chi2_contingency() for this test, calculated the confidence interval for 99% and compared the values for the hypotheses.
4. We used f_oneway() for wages of players with different international reputations and found the p-value.
5. Performed Wilcoxon Signed-rank Test for p-value, passing the difference of wages from 25000. Passed "method = 'approx'" to approximate and calculate p-value.

TAKEAWAYS AND CONCLUSIONS

Takeaways

1. 55% of all young adults have a BMI above 25.
2. There is a difference in BMI between public and private school students.
3. Most of the players joined from 2013 with the highest number of players joining in 2018. The number of new players has increased over the years.
4. The earliest player joined in 1991.
5. Most of the players have a release clause below 25,000,000. The highest release clause is more than 225,000,000.



6. Most of the players prefer their right foot.
7. Most of the players have a work rate of 'Medium/ Medium' or 'High/ Medium'. Least number of players have a work rate of 'Low/ Low'.
8. Most of the players have a normal or lean body.
9. Most of the players have an 'ST' position.
10. The values of the top 20 players increase with the overall scores of the players.
11. The overall rating of left-footed players is higher than that of right-footed players.
12. The potential of players aged greater than 35 is less than that of players aged less than 35.
13. The median wage of the top 20 players is more than 25000.

Conclusions

1. As the Data Science Team, to recommend 20 players for the main team of the club ‘Brussels United FC’ from the large data of players acquired through FIFA, we considered age, wage, work rate, preferred foot, body type, position, release clause, reputation, height and weight of all the players and how they affect their overall scores and ratings.
2. We also found out the top 5 players by overall rating for each unique position, finding the average wage one can expect to pay for them, as additional choices.

FUTURE STEPS

- A statistical test can be performed to determine the usual BMI of a young adult with a certain confidence level to have a standard of healthy young adult BMI for medical purposes.
- Instead of just the overall score, variables like ‘Potential’, ‘Crossing’, ‘Finishing’ and other performance measures can also be looked for while considering players for the team.
- The release clause and value of players can also be analysed to find best players possible in minimum budget.
- Advanced visualization tools such as Tableau or Power BI can be used to express our visual findings in detail.