MINI PROJECT PRESENTATION
EDA + STATS

EDA AND STATISTICAL ANALYSIS OF FIFA CASE STUDY

GROUP 3

Team Members: Aditya Aryan | Bhavani Kanaparthi | Naman Goswami | Neha Mondal

Introduction to the Problem Statement

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. We are given the task to use a data-based approach to 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.

Data Description

- Information about different players, their clubs and performance metrics
- 25490 entries
- 60 attributes
- 18 categorical attributes
- 42 numerical attributes
- 11833 duplicates
- 1171 maximum null count

	ID	Name	Age	Photo	Nationality	Flag	Overall
0	240331	P. Camará	21	https://cdn.sofifa.org/players/4/19/240331.png	Guinea Bissau	https://cdn.sofifa.org/flags/119.png	58
1	183465	J. Rodwell	27	https://cdn.sofifa.org/players/4/19/183465.png	England	https://cdn.sofifa.org/flags/14.png	68
2	205186	P. Gazzaniga	26	https://cdn.sofifa.org/players/4/19/205186.png	Argentina	https://cdn.sofifa.org/flags/52.png	74
3	233531	Y. Soteldo	21	https://cdn.sofifa.org/players/4/19/233531.png	Venezuela	https://cdn.sofifa.org/flags/61.png	71
4	243718	R. Koot	18	https://cdn.sofifa.org/players/4/19/243718.png	Netherlands	https://cdn.sofifa.org/flags/34.png	56

Data Cleaning Steps

1. Removed insignificant columns ‘Photo’, ‘Flag’, ‘Club’, ‘Club Logo’, ‘Jersey Number’ and ‘Loaned From’
2. Converted ‘Value’, ‘Wage’, ‘Joined’, ‘Contract Valid Until’, ‘Height’, ‘Weight’ and ‘Release Clause’
3. Dropped duplicates
4. Dropped <2% nulls
5. Imputed remaining
6. Dropped garbage values of ‘Body Type’

```
f['Body Type'].value_counts()
```

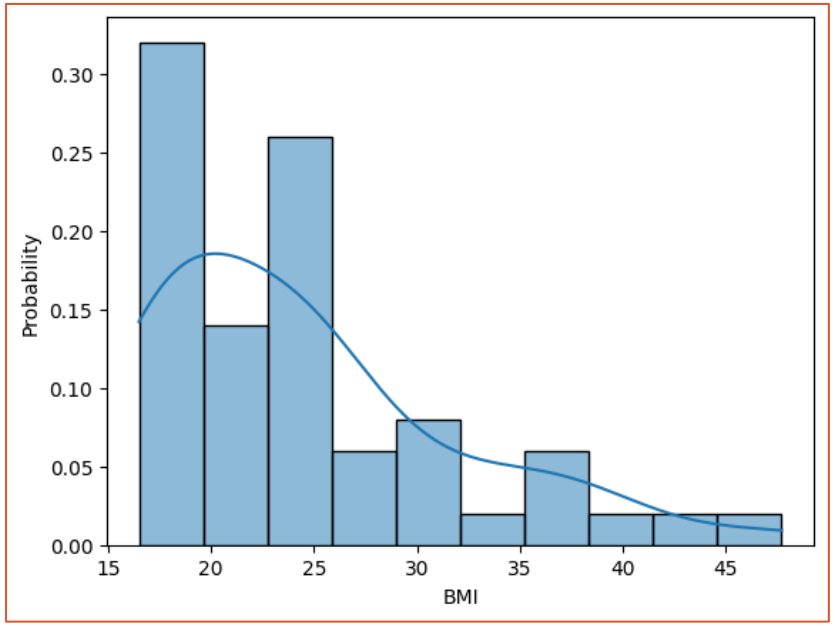
```
Normal                7826
Lean                  4769
Stocky                 835
Shaqiri                  1
C. Ronaldo               1
Neymar                   1
PLAYER_BODY_TYPE_25      1
Akinfenwa                1
Courtois                 1
Name: Body Type, dtype: int64
```

```
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()
```

Problem Solving Steps

Part – A

- Found basic stats of the BMI values
- Analysed outliers
- Found % rank and probability of BMI = 25 data point
- Plotted frequency and probability distribution
- Plotted KDE and sampling distributions
- Found probability of 6/10 adults having BMI > 25




```
n=1
plt.figure(figsize=(12,8))
random.seed(1) # sets random seed = 1 for the selection of same random values
for i in range(10,31,5): # iterates i from 10 to 30 with a step size of 5
    m=[np.mean(random.choices(b,k=i)) for j in range(100)]
    # takes 100 samples of sample size i with replacement, calculates their means and stores into m
    plt.subplot(2,3,n) # adds and selects nth subplot within a figure with 2 rows and 3 columns
    sns.kdeplot(x=m) # plots a Kernel Density Estimate (KDE) plot
    plt.xlabel('Sample Mean')
    plt.title('Sample size = '+str(i)) # sets title of subplot as 'Sample size = i'
    n+=1
plt.tight_layout() # adjusts spaces between subplots
plt.show()

# We can see that the sampling distributions with sample sizes 10 to 30 are less skewed than that with
# sample size 5. Also, the sampling distributions get less and less skewed as the sample size
# increases. This corroborates the Central Limit Theorem.
```

- Found probability of 50/100 adults having BMI > 25
- Computed 95% CI of true BMI
- Computed sample sizes to estimate proportion of adults with BMI > 25
- Tested claim of 55% adults having BMI > 25
- Tested claim of public and private school students having same BMI

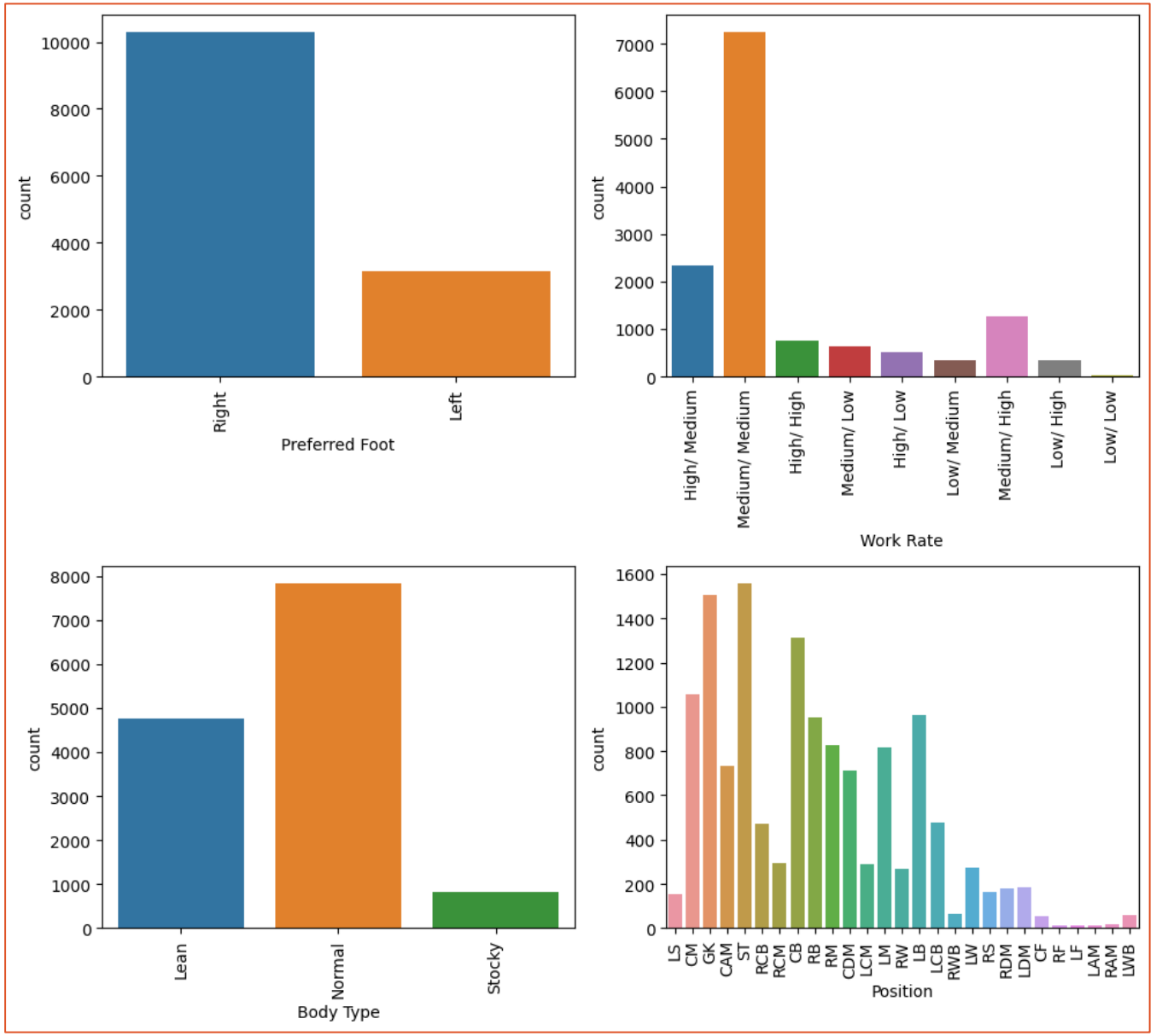
Q17. A data scientist wants to estimate with 95% confidence the proportion of young adults having BMI greater than 25.0. A recent study showed that 40% of all young adults have BMI greater than 25.0. The data scientist wants to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

```
z=1.96 # critical value for 95% confidence
p=0.4
e=0.02
n=ceil(z**2*p*(1-p)/e**2) # uses formula for sample size and then rounds to next nearest integer
n
```

```
2305
```

Part – B

- Checked variation of features
- Analysed outliers
- Checked for imbalance or cardinality in variables
- Created pair plot
- Listed top 20 players by score with contract end in 2020
- Computed their mean wage and age
- Analysed relation between their values and ratings
- Listed position-wise top 5 players by rating



- Computed position-wise mean wages
- Tested claim of left-footed players having higher rating
- Tested claim of players aged > 35 having lesser potential
- Tested relation between 'Preferred Foot' and 'Position'
- Tested effect of 'International Reputation' on 'Wage'
- Tested claim of top 20 players having median wage <= 25000

```
t=pd.Series(index=f.Position.unique(),dtype=float)
# creates a Series t with indices as unique position names and datatype floating number
for i in f.Position.unique():
    t.loc[i]=f[f.Position==i].nlargest(5,'Overall').Wage.mean()
    # stores mean of wages of top 5 players by overall score for position i into t
t
```

```
LS     114200.0
CM     122800.0
GK     138200.0
CAM    168000.0
ST     259000.0
RCB    219000.0
RCM    218600.0
CB     139600.0
RB     136000.0
RM     104000.0
CDM    179400.0
LCM     93200.0
LM     131200.0
RW     163200.0
```

```
# H0: 'Preferred Foot' and 'Position' are independent.
# H1: 'Preferred Foot' and 'Position' are not independent.

d=(len(f['Preferred Foot'])-1)*(len(f.Position)-1) # calculates degrees of freedom using formula
print('Test statistic =',stats.chi2_contingency(pd.crosstab(f['Preferred Foot'],f.Position))[0])
# performs Chi-square Test of Independence and returns test statistic
print('Confidence interval =',list(np.round(stats.chi2.interval(.99,d),2)))
# returns confidence interval for Chi-square distribution

# Since the test statistic does not lie within the confidence interval, we reject the null hypothesis.
# Hence, there is enough evidence to conclude that 'Preferred Foot' and 'Position' are not
# independent. Thus, they are related to each other.
```

```
Test statistic = 3319.754844426529
Confidence interval = [180289125.96, 180386963.55]
```

What could have been done better?

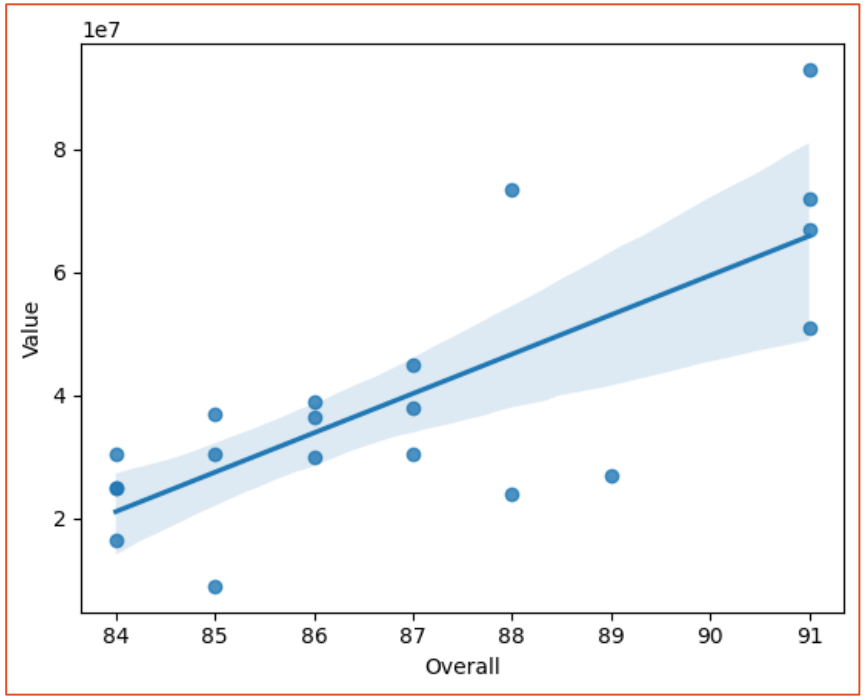
Imputation of the <2% nulls based on patterns in the data instead of dropping

```
# Since 'Contract Valid Until' has only 1.62 % missing values, we can drop all records with a null
# value in 'Contract Valid Until' column.

f=f[f['Contract Valid Until'].notna()]
# stores all records with a non-null value in 'Contract Valid Until' column into f
f.isna().sum()
```

Takeaways and Conclusions

- BMI of 55% young adults > 25
- Different BMIs of private and public school students
- Year of most joins: 2018
- Usual release clause < 25,000,000
- Most common position: ST
- Increase in values with scores of top players
- Higher rating of left-footed players
- Higher potential of players aged < 35
- Median wage of top players > 25000



Future Steps

- Statistical estimation of BMI of adults for medical purposes
- Looking into other performance measures as well instead of just score to recommend players
- Analysis of release clause and value to find players in minimum budget
- Usage of Tableau or Power BI to express findings

Thank you.