PGP Data Science Engineering
Exploratory Data Analysis – Mini Project:

A new football club named 'Brussels United FC' has just been inaugurated. This club does not have a team yet. The team is looking to hire players for their roster. Management wants to make such decisions using data based approach. During a recent hiring drive, you were selected for the Data Science team as a Junior data scientist. Your team has been tasked with creating a report which recommends players for the main team. To start with, a total 15 players are required. Player data for all teams has been acquired from FIFA. This data contains information about the palyers, the clubs they are currently playing for and various performance measures.

There is a limited budget for hiring players. The team needs 20 possible players to choose from. You have been requested to formulate a report in order to help the management make a decision regarding potential players.

Data:

The data contains details for over 18,000 players playing in various football clubs in Europe. It contains information on age, skill rating, wages and player value, etc. The files provided are as folows:
fifa.csv – data file.
fifa_ variable_information.csv - information on individual variables.

Data Preprocessing:

1. Import the necessary libraries and read the data.
2. Drop any columns that you deem unecessary for analysis.
3. The following columns need to be converted for further analysis:

Column	Details	Required output
'Value'	Amount with Euro symbol as prefix and suffix 'K' or 'M' indicating thousands and millions respectively.	Convert to Float after getting rid of currency symbol and suffix.
'Wage'	Amount with Euro symbol as prefix and suffix 'K' or 'M' indicating thousands and millions respectively.	Convert to Float after getting rid of currency symbol and suffix.
'Joined'	Year as a string, in some cases complete date as string	Convert to int with only year
'Contract Valid Until'	Date as a string	Convert to datetime type
'Height'	In inches with a quotation mark	Convert to Float with decimal points
'Weight'	Contains the suffix lbs	Remove the suffix and convert to float
'Release Clause'	Amount with Euro symbol as prefix and suffix 'K' or 'M' indicating thousands and millions respectively.	Convert to Float after getting rid of currency symbol and suffix.

(You might encounter Nan values in the above columns. Pandas treats Nan values as float. Please keep that in mind when making the conversions.)

4. Check for missing values and do a mean imputation where necessary.

Exploratory Analysis:

1. Plot the distribution of Overall rating for all players.
2. Generate pair plots for the following variables:
   Overall, Value, Wage, International Reputation, Height, Weight, Release Clause

3. Generate a table containing the top 20 players ranked by Overall score and whose contract expires in 2020.
    a) What would the average wage for this set of players be?
    b) What is the average age?
    c) Is there a correlation between the Overall rating and Value for these players?

4. Generate tables containing the top 5 players by Overall rating for each unique position.
    a) Are there any players appearing in more than one Table. Please point out such players.
    b) What is the average wage one can expect to pay for the top 5 in every position?

Final Report:
Put all highlights from the information obtained above in a power point presentation containing a maximum of 5 slides. No title slide required.