# Data Visualization on NBA Dataset

## Data Visualization

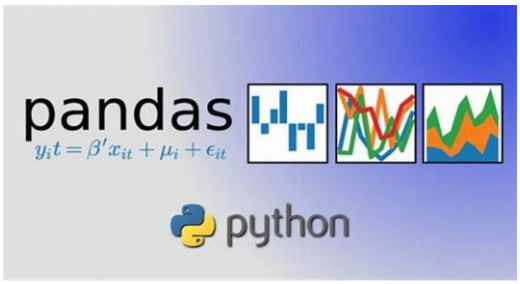
- Data Visualization is the presentation of data in pictorial format.
- Helps in better utilization of data.

## Python

- Python is widely used multi-purpose high level programming language.
- Broad standard library -It has rich libraries of pre-defined functions for numerous applications like data visualization.

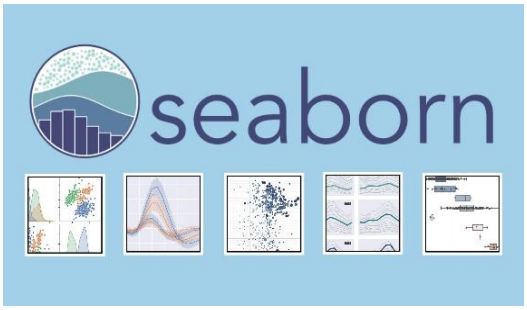
Libraries used-

## Pandas-




- Pandas is an open-source library that is used for data analysis.
- It allows efficient data cleaning and manipulation operations.

## Seaborn-




- Seaborn is built on top of Python's core visualization library Matplotlib.
- Seaborn works well with Num Py and Pandas data structures.

## Dataset Used-

- Nba csv file which contains data of Nba players.
- It contains the following dimensions-[457 rows X 9 columns].
- The link/reference to this dataset is- https://github.com/sivabalanb/Data-Analysis-with-Pandas-and-Python/blob/master/nba.csv

## Platform used-

- Google colab for python coding.

**Madhurima Rawat(Datascience Csvtu)**

## Steps involved are:-

- First CSV file is imported in Google Colab(Platform used for execution of python code).
- Then data visualization is done by using python library seaborn(as described above).

## Codes and Graphs along with description are as follows:

### Importing libraries and dataset

```
#Importing Pandas(Python library which will be used to import csv file)
 as pd
import pandas as pd
#Importing python library seaborn(Library to do data visualiztion)
import seaborn as sns
#Importing CSV file or dataset as dataframe
data=pd.read_csv('/content/drive/MyDrive/Nba.csv')
#printing Csv file
print(df)
```

### Output is as follows:

```
              Name            Team  Number Position  Age Height  Weight  \
0    Avery Bradley  Boston Celtics       0       PG   25    6-2     180
1      Jae Crowder  Boston Celtics      99       SF   25    6-6     235
2     John Holland  Boston Celtics      30       SG   27    6-5     205
3      R.J. Hunter  Boston Celtics      28       SG   22    6-5     185
4    Jonas Jerebko  Boston Celtics       8       PF   29   6-10     231
..             ...             ...     ...      ...  ...    ...     ...
452     Trey Lyles       Utah Jazz      41       PF   20   6-10     234
453   Shelvin Mack       Utah Jazz       8       PG   26    6-3     203
454      Raul Neto       Utah Jazz      25       PG   24    6-1     179
455   Tibor Pleiss       Utah Jazz      21        C   26    7-3     256
456    Jeff Withey       Utah Jazz      24        C   26    7-0     231

               College     Salary
0                Texas  7730337.0
1            Marquette  6796117.0
2    Boston University        NaN
3         Georgia State  1148640.0
4                  NaN  5000000.0
..                 ...        ...
452           Kentucky  2239800.0
453             Butler  2433333.0
454                NaN   900000.0
455                NaN  2900000.0
456             Kansas   947276.0

[457 rows x 9 columns]
```

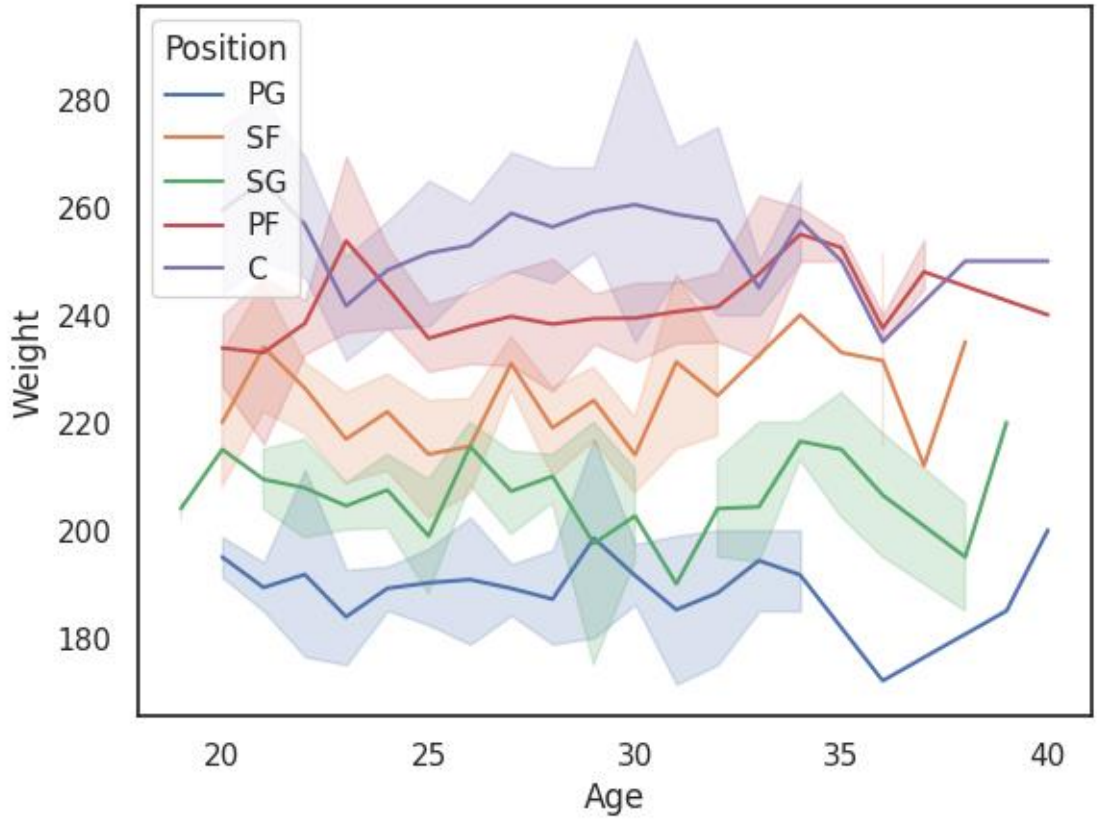
## Data visualization in the form of various form of graphs is as follows:

# Line Plot

Line plot is the most popular plot to draw a relationship between x and y with the possibility of several semantic groupings.

```
sns.lineplot(x='Age',y='Weight',data=df,hue=data["Position"])
```

## Output is as follows:




## Description of the Output/Graph

- Seaborn Line Plot helps to visualize the statistical relationships.
- Lineplot can be extensively used in situations wherein we feel the need to check the dependency of a parameter on the other in a continuous manner relative to time.
- To understand how variables in a dataset are related to one another and how that relationship is dependent on other variables, we perform statistical analysis.
- This Statistical analysis helps to visualize the trends and identify various patterns in the dataset.

**Madhurima Rawat(Datascience Csvtu)**

- In the above dataset ,we are plotting the line plot between two parameters weight and age.
- It also shows the semantic grouping with respect to the Position parameter of the dataset.

**Important conclusions that can be made from the above Graph:**

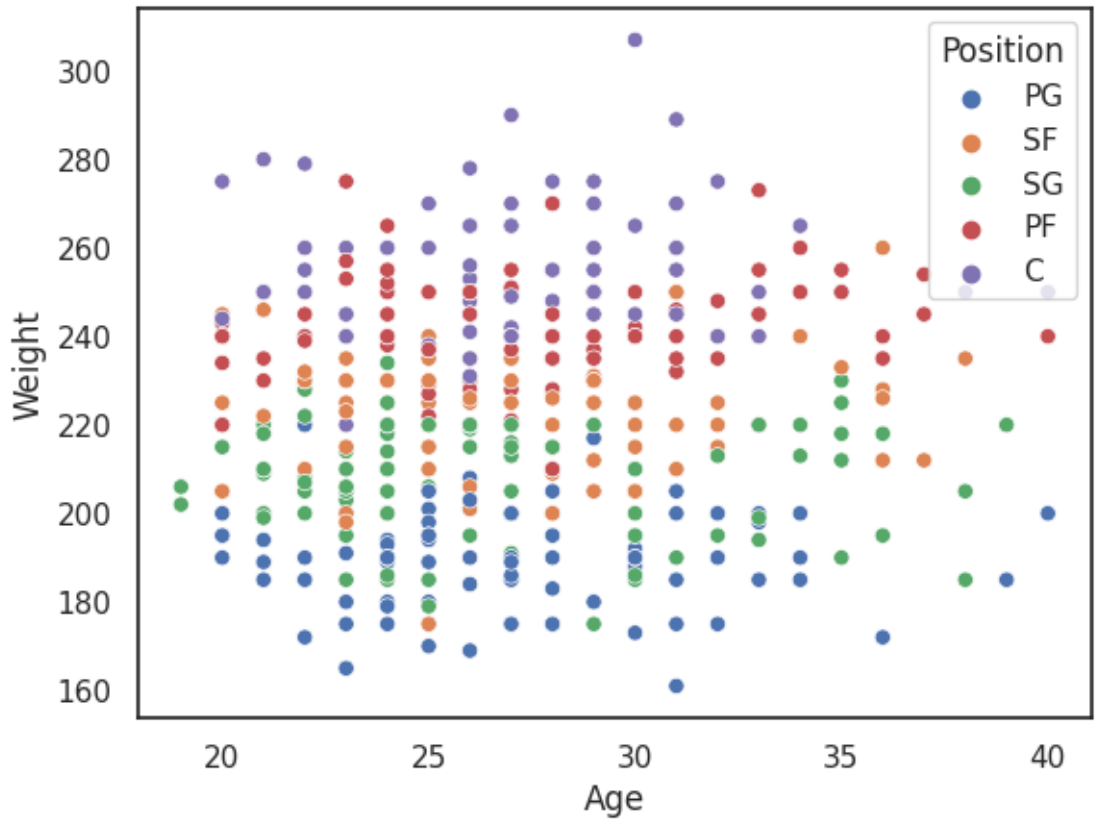
- PG position players have the highest weight age correspondence.
- The graph also shows each value of Age and Weight along with sematic grouping with respect to Position parameter which makes it easier to understand the two parameters with respect to each other.

## Scatter Plot

- Scatterplot Can be used with several semantic groupings which can help to understand well in a graph against continuous/categorical data.
- It can draw a two-dimensional graph.

```
sns.scatterplot(x='Age',y='Weight',data=df,hue=data["
Position"])
```

**Output is as follows:**



## Description of the Output/Graph

- Statistical Analysis is the basic estimation out of some parameters of the dataset to a large extent.
- Data Visualization is considered to be the best way to perform statistical analysis i.e. predict the outcome or the cause based on diagrammatic values.
- The seaborn scatterplot is basically used to depict the relationship between the parameters on the given axes respectively.
- Every point on the graph depicts a value corresponding to it.
- In the above dataset ,we are plotting the scatter plot between two parameters weight and age.
- It also shows the semantic grouping with respect to the Position parameter of the dataset.
- Now in the above graph, we can see that each point represents the relationship or values of age and weight of Nba players along with representing the groupings of the semantic parameter "Position".
- Now the difference between this scatter and previous line plot is that in the line plot we need the values to have some relationship between them but in scatter plot ,the graph can be drawn without this as well.
- Also scatter plot is more suited to data with many outliers or variation values.

**Important conclusions that can be made from the above Graph:**

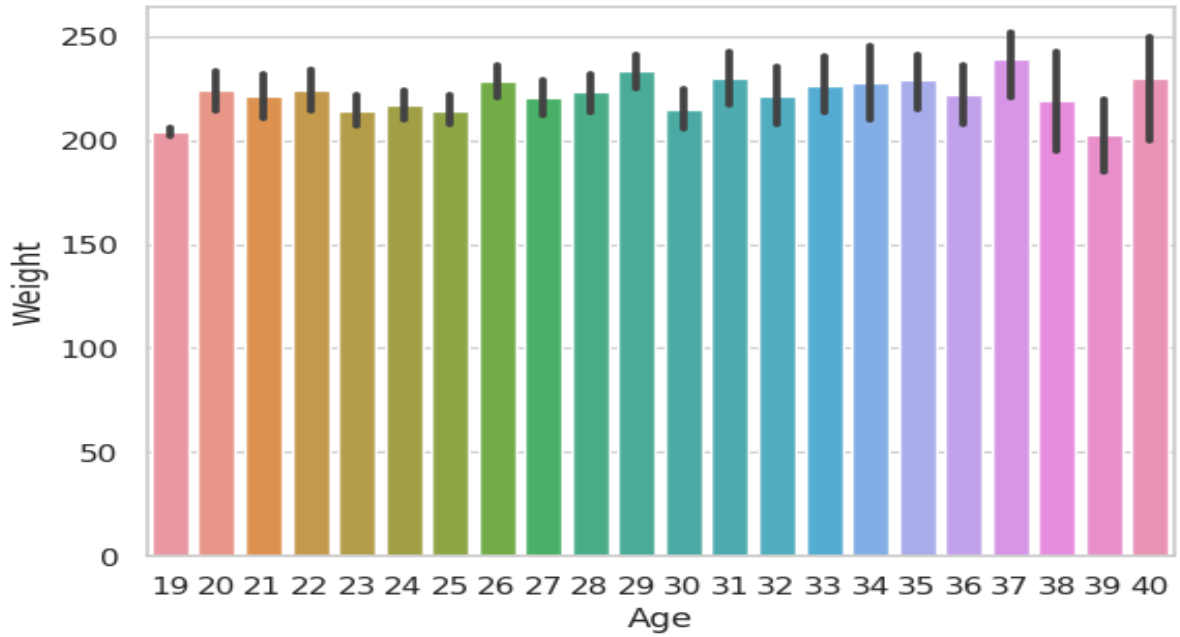
- PG position players have the highest weight age correspondence.
- The graph also shows each value of Age and Weight along with sematic grouping with respect to Position parameter which makes it easier to understand the two parameters with respect to each other.
- Also since the values are plotted with points instead of lines,it is very easy to find the exact value for any particular age.
- It could be easily seen that if we want to calculate any data related to age,position and weight ,with these graph just by looking and calculating the points , calculations can be easily made.

## Bar Plot

- Barplot represents an estimate of central tendency for a numeric variable.

```
sns.barplot(x='Age',y='Weight',data=df)
```

**Output is as follows:**




## Description of the Output/Graph

- Bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars.
- Categorical estimate plots-The estimation of categorical data basically refers to the representation of certain estimation or prediction of the categorical data values to the corresponding data variable.
- Python Seaborn has the Bar plot to be used for the estimation of categorical data.
- Now in the above dataset ,every rectangle on the graph depicts a value corresponding to it.
- In the above dataset ,we are plotting the bar plot between two parameters weight and age.
- Now in the above graph, we can see that each rectangle represents the relationship or values of age and weight of Nba players.

## Important conclusions that can be made from the above Graph:

- Easier to find any value of Weight corresponding to Age.
- Easier to find the highest/lowest value of Age corresponding to weight and vice versa.
- This method is accepting the parameters Age, Weight and hue it is an optional parameter it helps to take column name for color encoding.

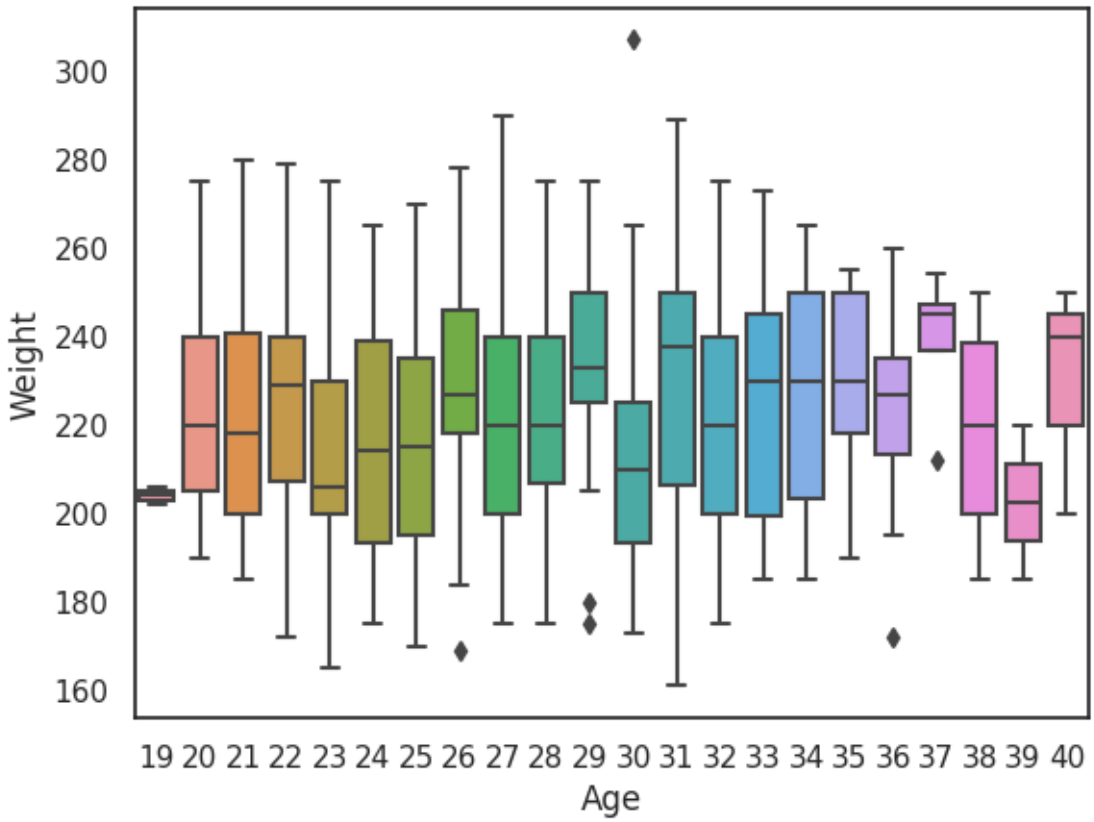
# Box Plot

- A box plot (or box-and-whisker plot) is the visual representation of the depicting groups of numerical data through their quartiles against continuous/categorical data.

Code is as follows:

```
sns.boxplot(x='Age',y='Weight',data=df)
```

**Output is as follows:**




## Description of the Output/Graph

- A box plot (or box-and-whisker plot) is the visual representation of the depicting groups of numerical data through their quartiles against continuous/categorical data.
- A box plot consists of 5 things.
    1. Minimum
    2. First Quartile or 25%
    3. Median (Second Quartile) or 50%
    4. Third Quartile or 75%

5. Maximum

- Now in the above dataset ,every box on the graph depicts a value corresponding to it.
- In the above dataset ,we are plotting the box plot between two parameters weight and age.
- Now in the above graph, we can see that each box represents the relationship or values of age and weight of Nba players.

## Important conclusions that can be made from the above Graph:

- The box plot represents the categorical distribution of data and sets comparison among the different categorical data inputs.
- The 'box' structure represents the main quartile of the data input while the 'line' structure represents the rest of the distribution of data.
- The outliers are represented by points using an inter-quartile function.

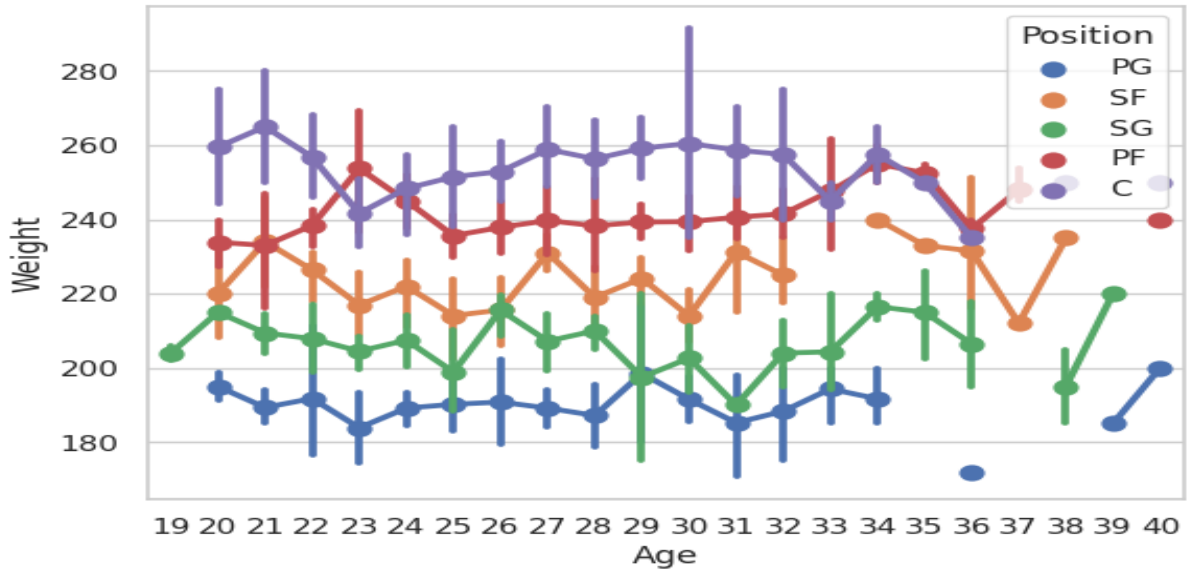
# Point Plot

- Point plot used to show point estimates and confidence intervals using scatter plot graphs.
- A point plot represents an estimate of central tendency for a numeric variable.

**Code is as follows:**

```
sns.pointplot(x = "Age", y = "Weight", data = df, hue=data["Position"])
```

**Output is as follows:**



## Description of the Output/Graph

- The Point Plot represents the estimation of the central tendency of the distribution with the help of scatter points and lines joining them.
- A point plot represents an estimate of central tendency for a numeric variable by the position of scatter plot points and provides some indication of the uncertainty around that estimate using error bars.
- In the above dataset, we are plotting the Point plot between two parameters weight and age.
- It also shows the semantic grouping with respect to the Position parameter of the dataset.
- Now in the above graph, we can see that each point represents the relationship or values of age and weight of Nba players along with representing the groupings of the semantic parameter "Position".
- Now the difference between this and previous plots is that in the previous plots we need the values to have some relationship between them but in point plot, the graph can be drawn without this as well.

## Important conclusions that can be made from the above Graph:

- The estimation of categorical data basically refers to the representation of certain estimation or prediction of the categorical data values to the corresponding data variable.
- PG position players have the highest weight age correspondence.
- The graph also shows each value of Age and Weight along with sematic grouping with respect to Position parameter which makes it easier to understand the two parameters with respect to each other.
- Also since the values are plotted with points instead of lines,it is very easy to find the exact value for any particular age.
- It could be easily seen that if we want to calculate any data related to age,position and weight ,with these graph just by looking and calculating the points , calculations can be easily made.

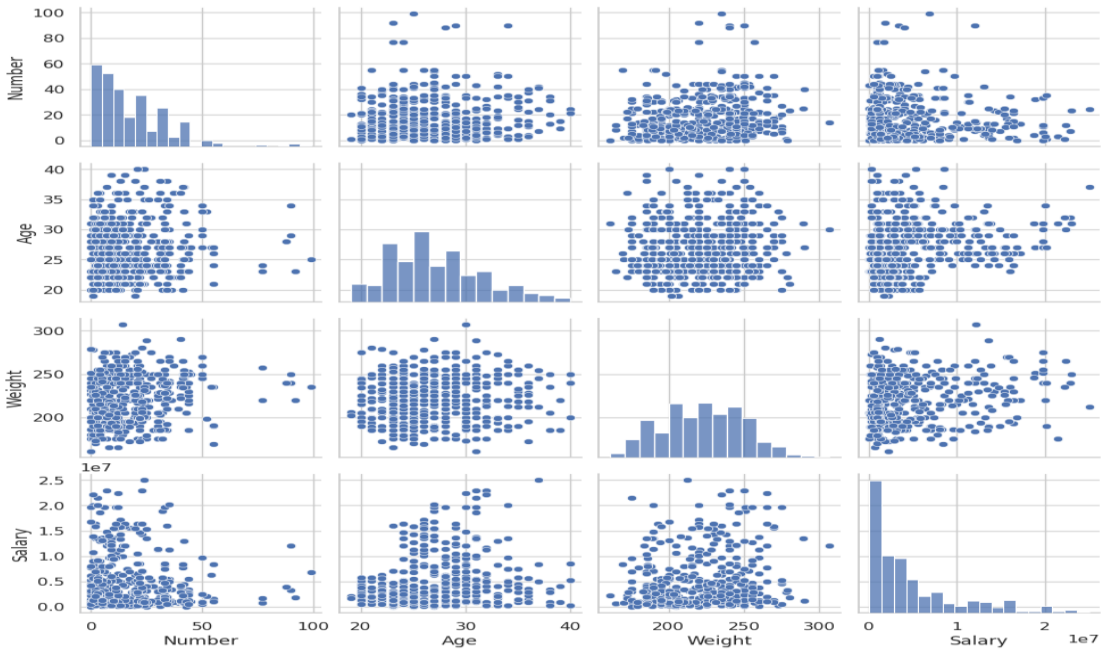
# Pair Plot

- It takes all the numerical attributes of the data and plot pairwise.

**Code is as follows:**

```
sns.pairplot(df)
```

**Output is as follows:**




## Description of the Output/Graph

- A pair plot is a data visualization that plots pair-wise relationships between all the variables of a dataset.
- This allows you to better understand the relationships visually, while even layering in additional details (such as by using color).
- Each variable is plotted both in the rows and columns, showing the relationships between the variables.
- By default, Seaborn will plot all of the different numeric variables in the dataset.
- Dataset Nba is plotted here and we can see that it plots all numerical Value column with respect to each other and also with the entire dataframe.

## Important conclusions that can be made from the above Graph:

- Pair plot creates a grid of axis such that each numeric variable in data will create a plot between each other the y-axis across a single row and the x-axis across a single column.
- The diagonal plots are a univariate distribution plot that helps to draw the marginal distribution of the data in each column.
- A pair plot pairwise relationships with other columns in the data frame and also plot pair plot with itself.
- **With these Graph it is very easy to compare values in an entire dataframe as it plots all values in one Graph.**
- This is why pair plots are one of the best data representations in seaborn.