SDS Assignment – Olympics Data Set 6.csv

Name: Atharva S Gadad

SRN: PES1UG20CS088

Sec: B

Data Cleaning

- Found that there were NULL values only in numerical columns such as Age, Height and Weight
- Replaced the Null values with mean values of the respective columns as instructed

Distribution of silver medalists

- Creates a new data frame with rows of silver medalists from the original data
- Plotted a histogram
- We find that the ages are almost normally distributed
- We find the highest concentration of silver medalists approximately between the ages 22 and 32
- This might be because this age frame has the peak of physical fitness

BMI

- We first converted the height in the data which was given in centimeters to meters
- Created a new column called BMI which was found by dividing weight in kgs by the height in meters squeared
- Found the BMI of players which are generally in the higher side owing to their high strength and muscular build

Height v/s weight

- Generated a scatter plot of height v/s weight
- Mostly very randomly scattered so hard to figure out correlation
- Based on the trends, assumed a very weak positive correlation
- Assumption was confirmed when Found the correlation coefficient to be around 0.56
- This can be attributed to the fact that different sports require different physiques – some require tall and lanky players, other want very well built and stout ones

Gender Distribution

- Found Gender distribution in the past 5 years for athletics which was the only sport in my dataset
- Found an approximately 50-50 split in the gender distribution
- This can be attributed to the fact that that for almost every male athletics even, there is a female athletics even as well
- Each event has 3 medallists and our dataset contains only data about medallists
- This implies that there are almost equal number of males and females in the data

TASK 1

- Found Mexico to have been the most participating country in the dataset
- Found their medal distribution and plotted stacked bar graph
- Found that they have won almost the same number of medals in all years except 2002 where their medal count was low
- They mostly win bronze medals, fewer silver medals and even fewer gold medals

TASK 2

- Generated a new data frame for athletes
- Found the number of medals and number of participation
- Found the Success Ratio
- Found the success ratio is 1 for most players
- Found the most successful player to have been Frentorish Torie Bowie with success ration of 3.0