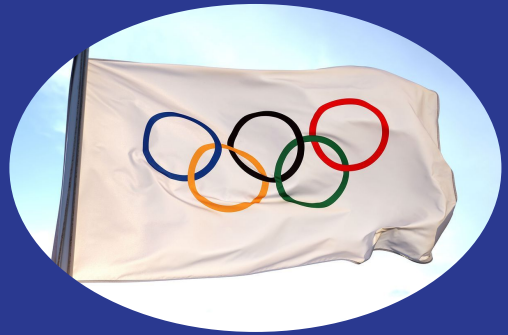
PR1 (Chapter 3): Descriptive Analysis for Olympic Dataset

CEN 4930 - CRN 85705 - TR - 202308

Andy, Chris, and Marcelo



The Dataset - Olympic Data

Rows

- Each row is represents an athlete.
- The dataset has 70k rows, that being data from 70k olympians over the years.

Columns

- The dataset has 15 columns / variables in total.
- Each variable brings details about the olympian, such as gender and NOC.

Availability

- The Dataset is available on Kaggle.
- The dataset has NaN / missing values

The Dataset - Overview

Link to dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data

BHANUPRATAP BISWAS · UPDATED 2 MONTHS AGO

100 | New Notebook | Download (1 MB)

Olympic Data

The Olympic Games are an international multi-sport event held every four years

Data Card | Code (8) | Discussion (1)

About Dataset

The Olympic Games are an international multi-sport event held every four years in which thousands of athletes from around the world participate in various sports competitions. The Olympics are one of the most significant and prestigious sporting events globally, promoting unity, friendship, and fair play among nations.

Key facts about the Olympic Games:

Usability
10.00

License
ODC Public Domain Dedication ...

Expected update frequency
Annually

The Dataset - Overview

Link to dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data

```
Data columns (total 15 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   ID      70000 non-null  int64
 1   Name    70000 non-null  object
 2   Sex     70000 non-null  object
 3   Age     67268 non-null  float64
 4   Height  53746 non-null  float64
 5   Weight  52899 non-null  float64
 6   Team    70000 non-null  object
 7   NOC     70000 non-null  object
 8   Games   70000 non-null  object
 9   Year    70000 non-null  int64
 10  Season  70000 non-null  object
 11  City    70000 non-null  object
 12  Sport   70000 non-null  object
 13  Event   70000 non-null  object
 14  Medal   9690 non-null   object
```

	ID	Age	Height	Weight	Year
count	70000.000000	67268.000000	53746.000000	52899.000000	70000.000000
mean	18081.846986	25.644645	175.505303	70.900216	1977.766457
std	10235.613253	6.485239	10.384203	14.217489	30.103306
min	1.000000	11.000000	127.000000	25.000000	1896.000000
25%	9325.750000	21.000000	168.000000	61.000000	1960.000000
50%	18032.000000	25.000000	175.000000	70.000000	1984.000000
75%	26978.000000	28.000000	183.000000	79.000000	2002.000000
max	35658.000000	88.000000	223.000000	214.000000	2016.000000

The Dataset - Overview

Link to dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data

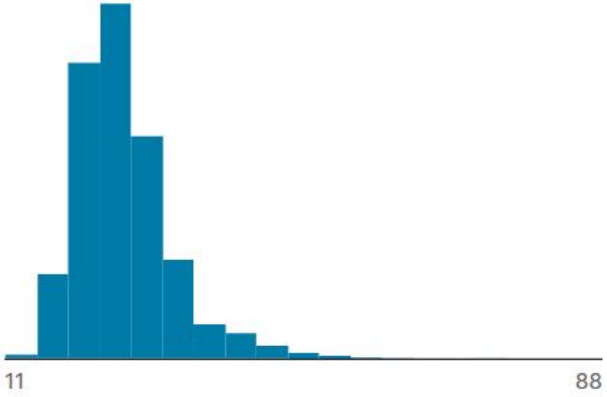
	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres

The Dataset - Overview

Link to dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data

Age

AG



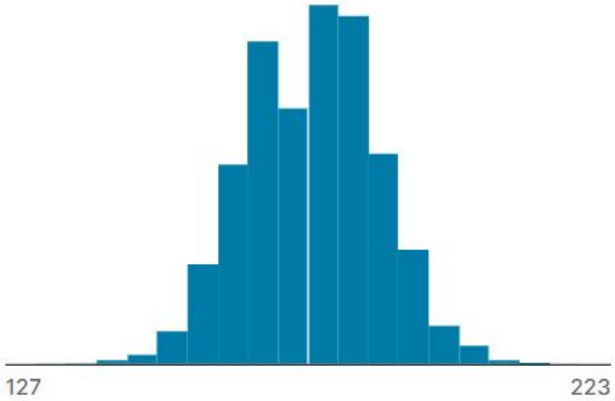
Valid	67.3k	96%
Mismatched	0	0%
Missing	2732	4%
Mean	25.6	
Std. Deviation	6.49	
Quantiles	11	Min
	21	25%
	25	50%
	28	75%
	88	Max

The Dataset - Overview

Link to dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data

 # Height

HG



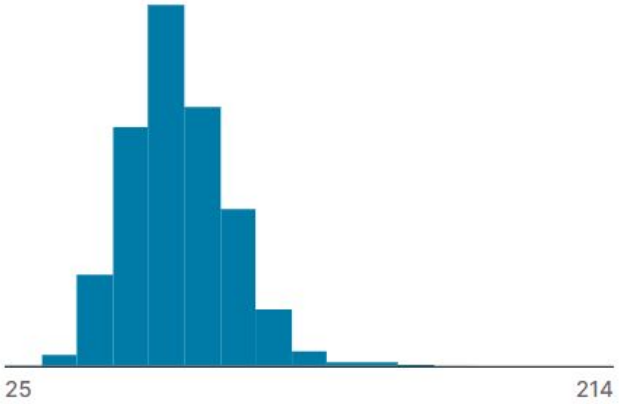
Valid	53.7k	77%
Mismatched	0	0%
Missing	16.3k	23%
Mean	176	
Std. Deviation	10.4	
Quantiles	127	Min
	168	25%
	175	50%
	183	75%
	223	Max

The Dataset - Overview

Link to dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data

 # Weight

WE



Valid	52.9k	76%
Mismatched	0	0%
Missing	17.1k	24%
Mean	70.9	
Std. Deviation	14.2	
Quantiles	25	Min
	61	25%
	70	50%
	79	75%
	214	Max

Challenges

Central Tendency and Dispersion measures

Central Values and Variations

The goal here is to analyze Central Values (Mean, Median and Mode) as well as understand the distribution of data.

Measures of Shape

Data Shape Visualization

By using graphing plots techniques, our goal is to visualize the impact of data manipulation into measures of shape.

Association Measures

Attributes Relationship

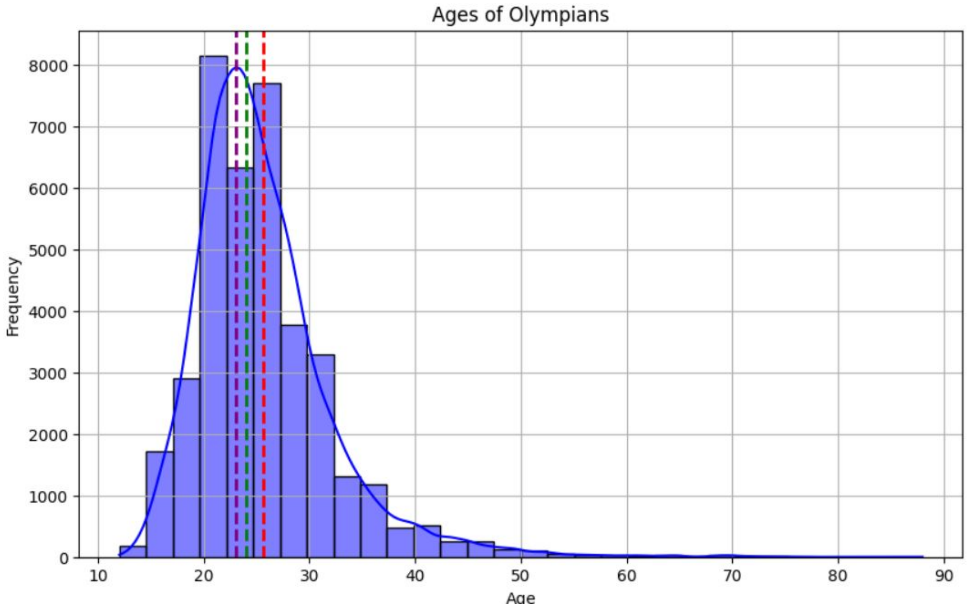
Numerically quantifying the relationships between the attributes.

Central Tendency Measures

The typical or Central Values

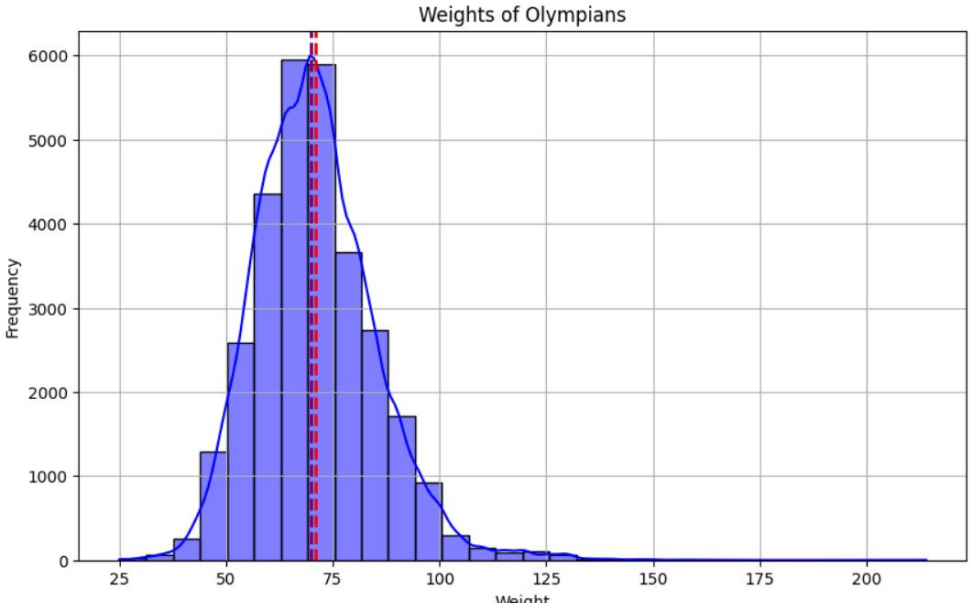
- Columns used:
  - Weight
  - Age
  - Height
- Measures:
  - Mean
  - Median
  - Mode

Age Measures of Central Tendency



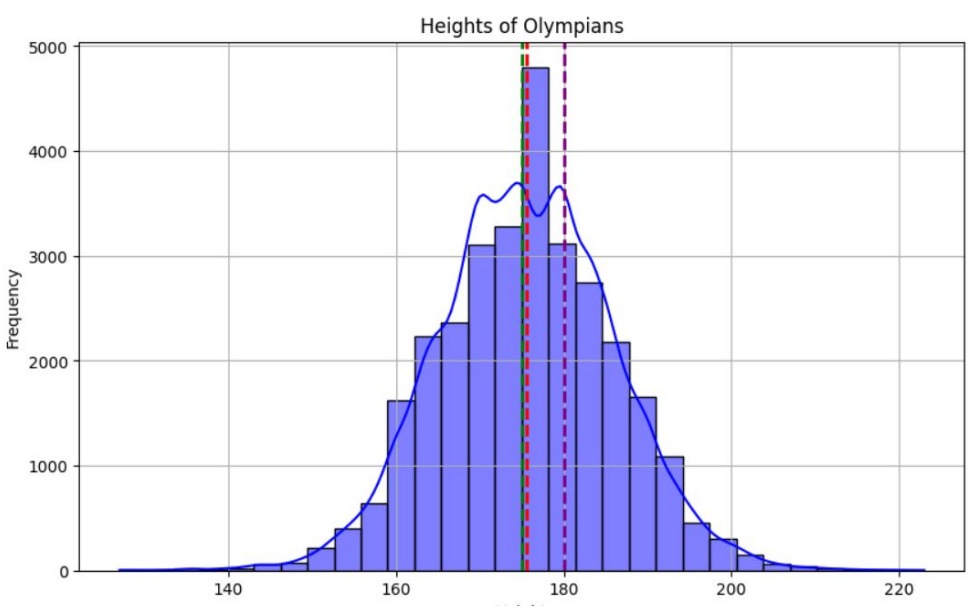
- Mean
  - 25.634768822933182
- Median
  - 25.0
- Mode
  - 23.0

Weight Measures of Central Tendency



- Mean
  - 70.9287878787
- Median
  - 70.0
- Mode
  - 70.0

Height Measures of Central Tendency



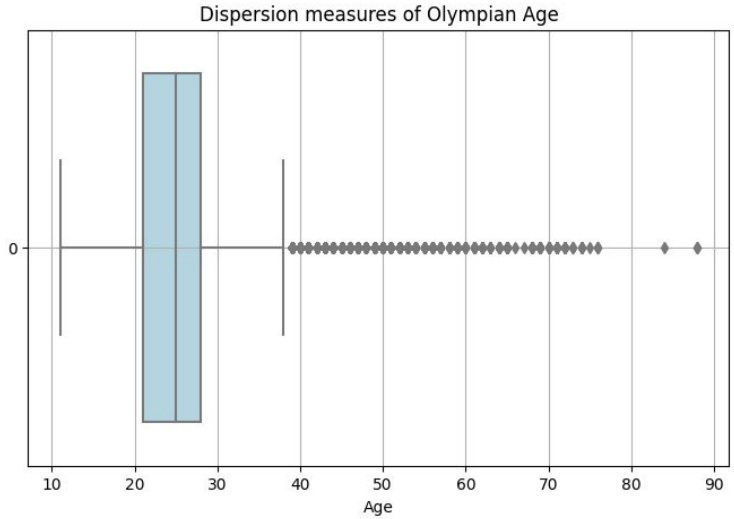
- Mean
  - 175.50353132628152
- Median
  - 175.0
- Mode
  - 180.0

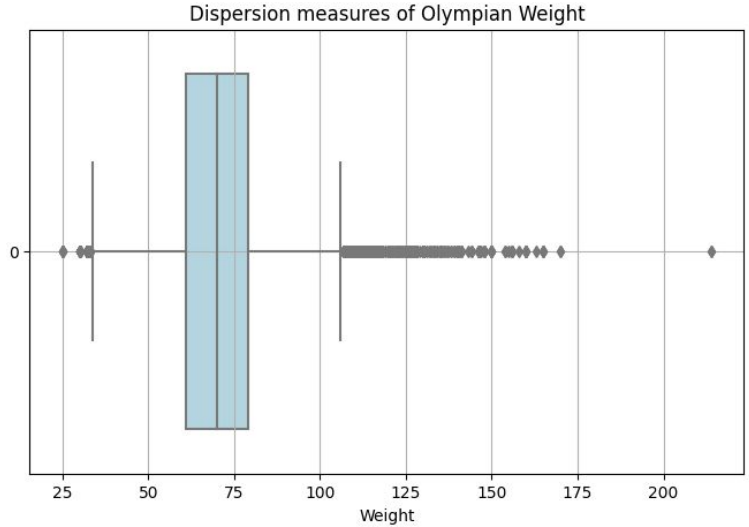
Dispersion Measures

Understanding the distribution of the data

- Columns used: Age, Weight, Height
- Methods: Range, IQR, Variance, Standard Deviation

Dispersion

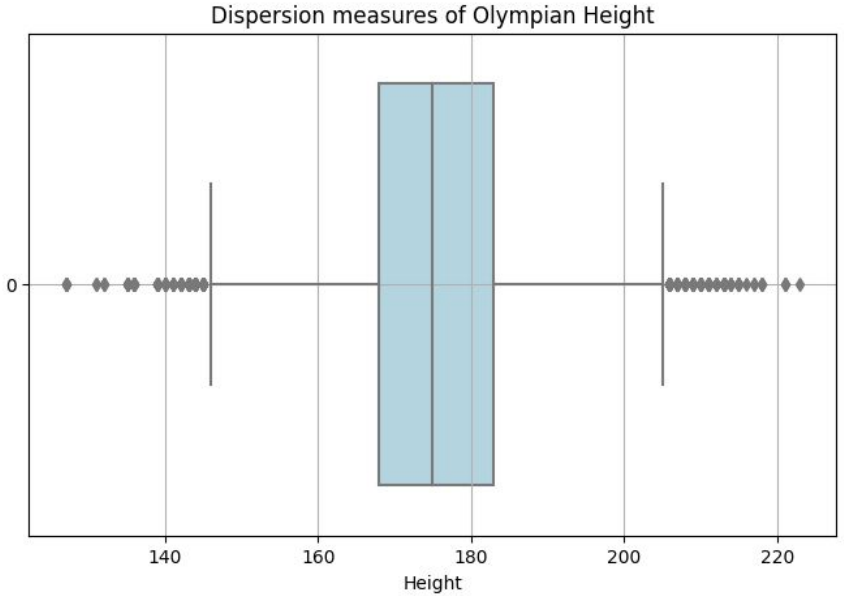





```
Range: 77.0
IQR: 7.0
Variance: 42.05832876799684
Standard Dev.: 6.485239299208383
```

```
Range: 189.0
IQR: 18.0
Variance: 202.13699706749927
Standard Dev.: 14.217489126688273
```

Dispersion cont.



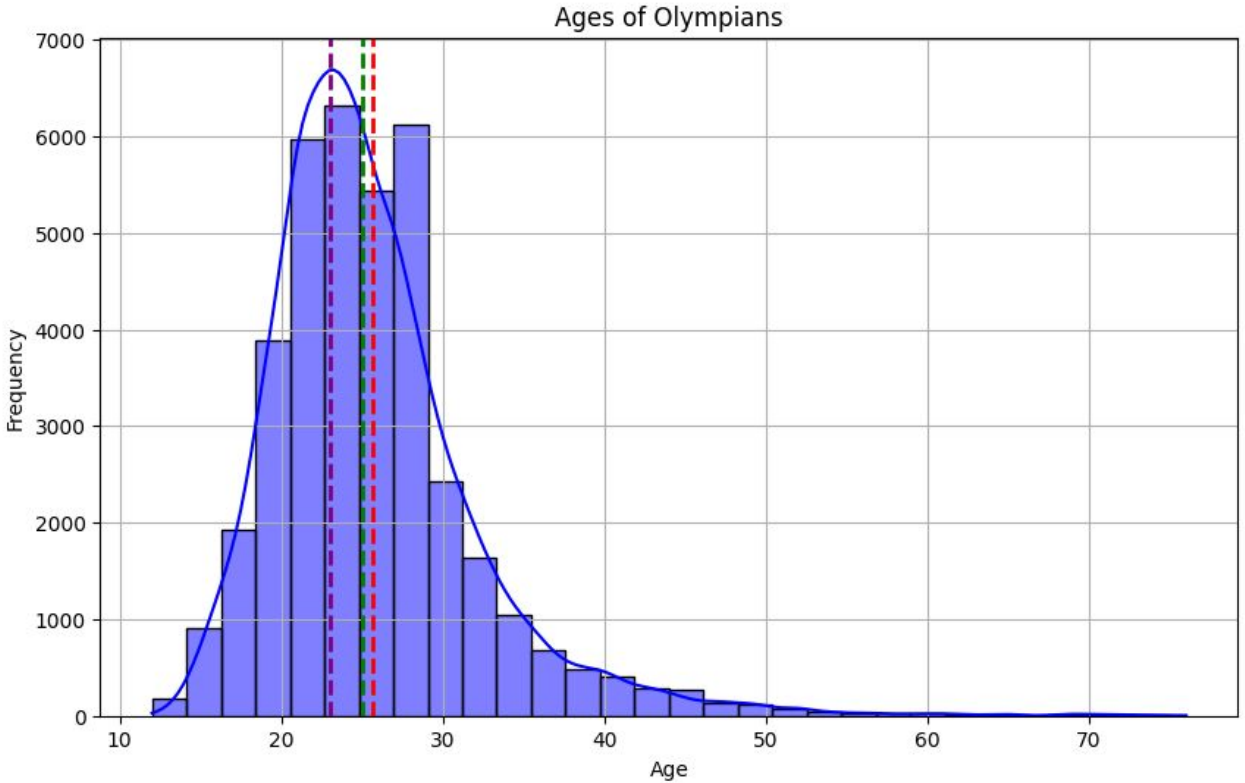

```
Range: 96.0
IQR: 15.0
Variance: 107.83166785235356
Standard Dev.: 10.384202802928762
```

Measures of Shape

Analysis into shape impact

- Columns used: Age, Weight, Height
- Methods: Kurtosis, Skewness

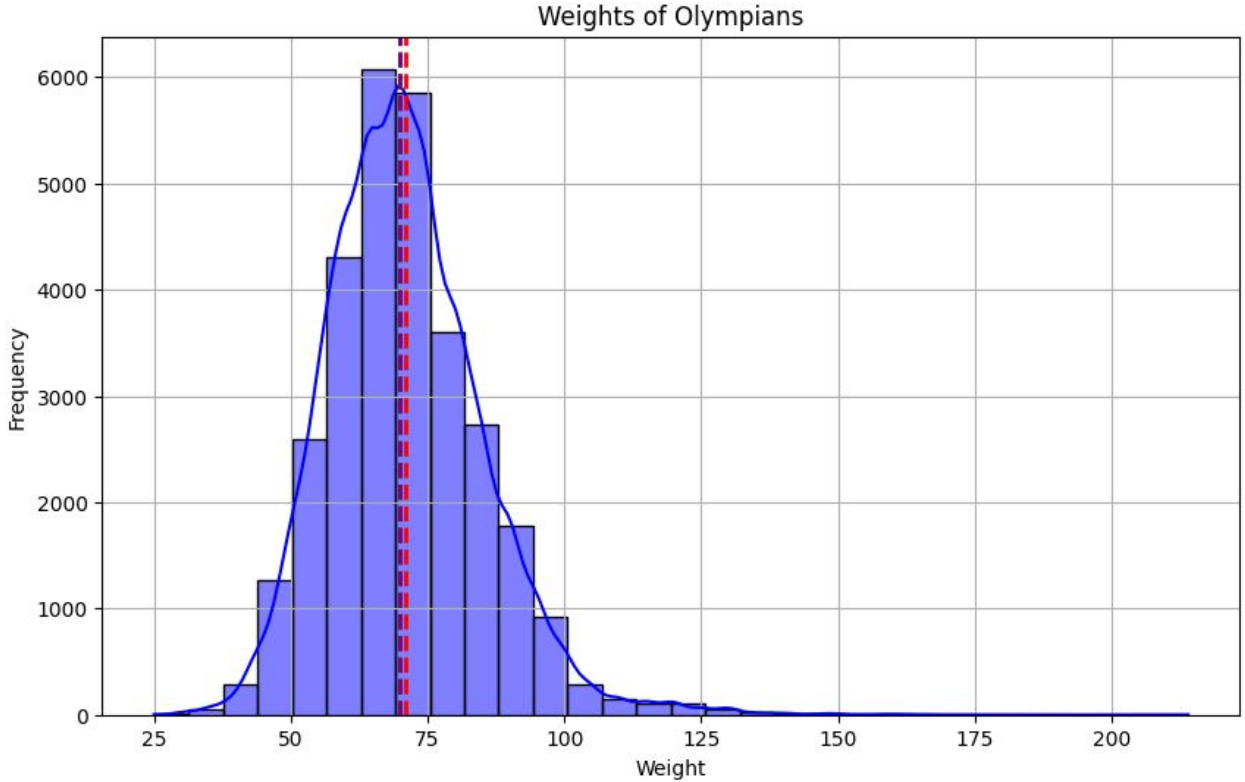
Measures of Shape - Age



Skewness: 1.717252
Kurtosis: 6.053829

Right-skew
Leptokurtic distribution

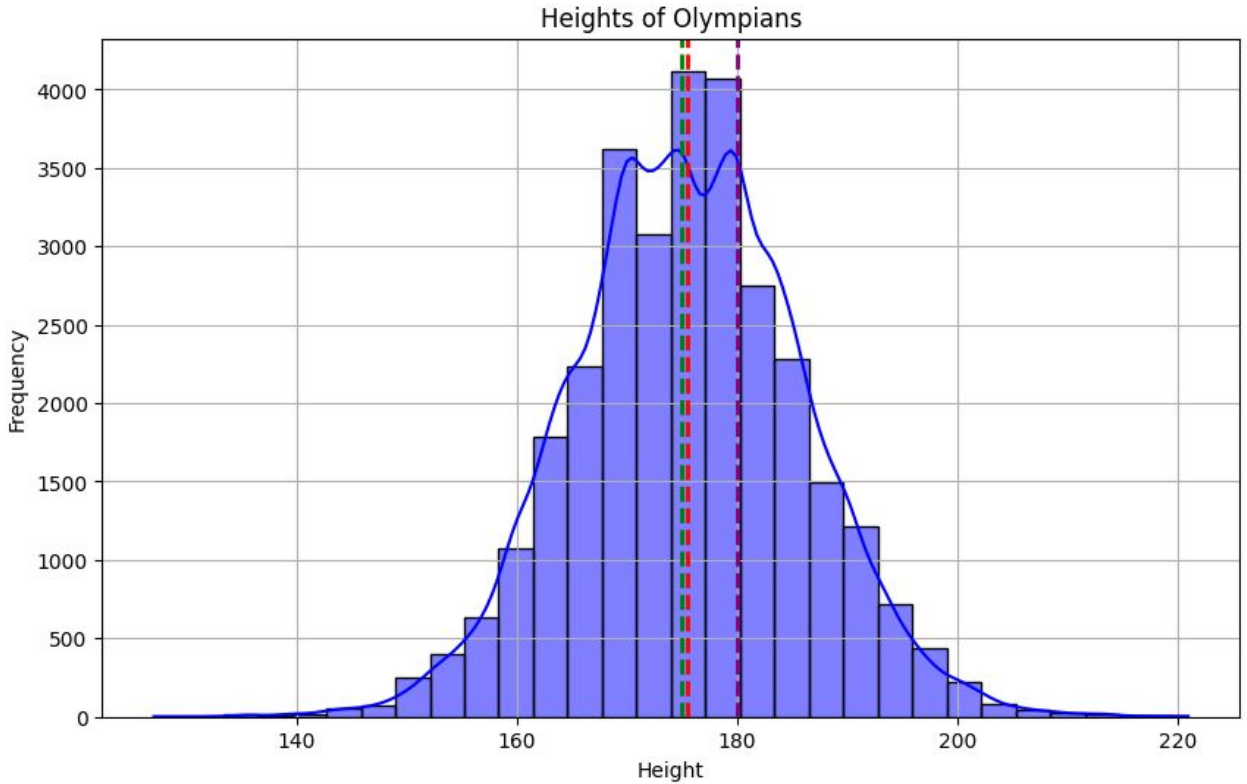
Measures of Shape - Weight



Skewness: 0.821064
Kurtosis: 2.326785

Very slight right-skew
Leptokurtic distribution

Measures of Shape - Height



Skewness: 0.017584
Kurtosis: 0.233221

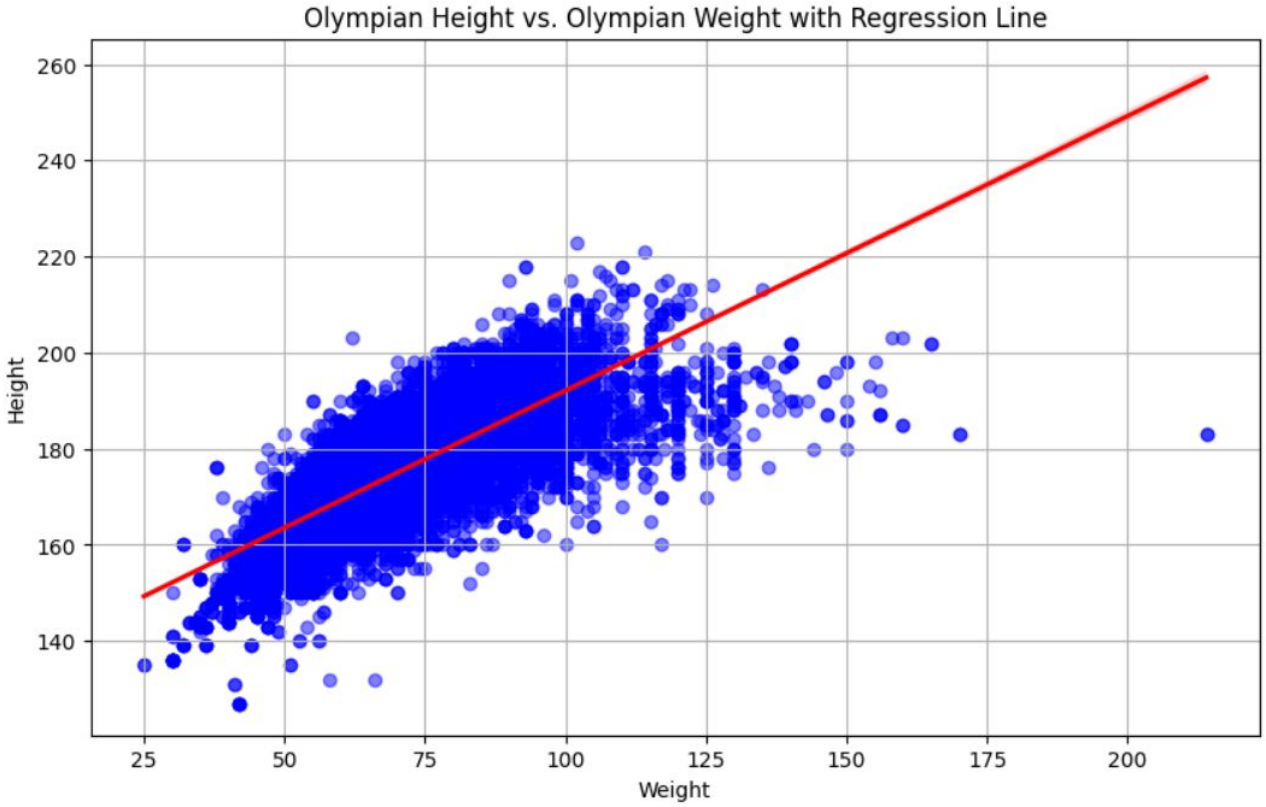
No skew
Mesokurtic distribution

Association Measures

Quantifying relationships between attributes

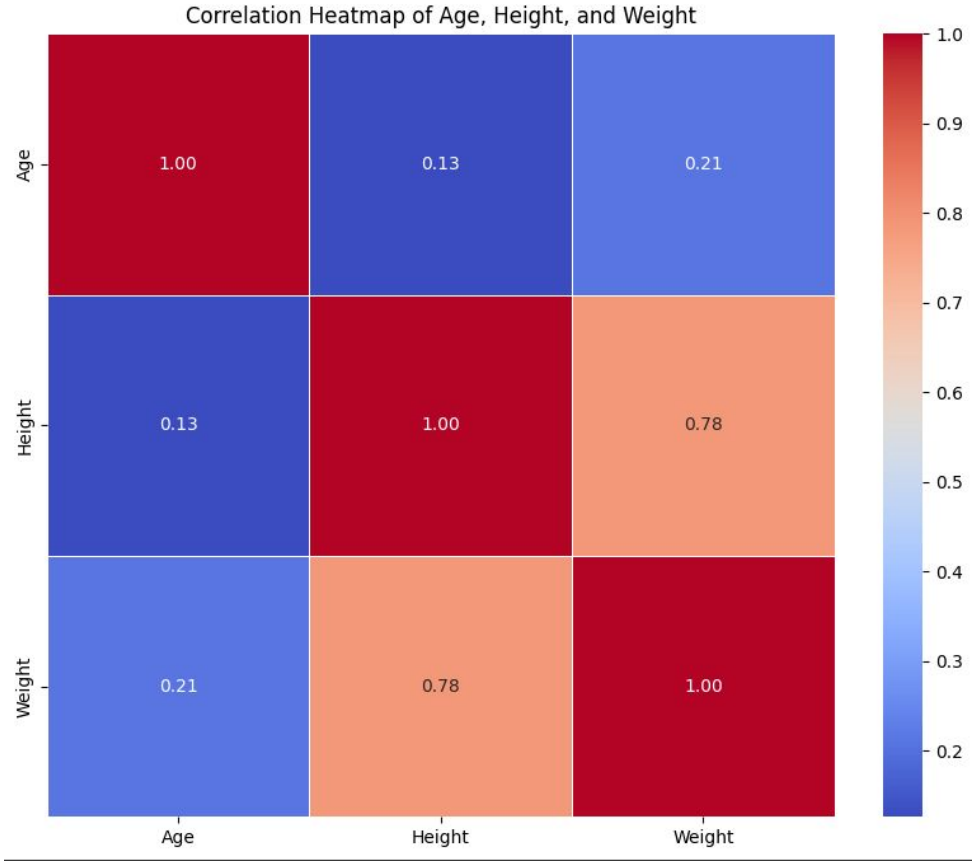
- Columns used:
  - Weight
  - Age
- Measures:
  - Correlation
  - Covariance

Relationship Between Height and Weight



- Covariance: 115.77
- Correlation: 0.78

Heatmap


Correlation Heatmap of Age, Height, and Weight
Age
Height
Weight
1.00
0.13
0.21
0.13
1.00
0.78
0.21
0.78
1.00
Age
Height
Weight
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Results Obtained

Results Obtained

- The **Age Graphic** displays a **right-skewed** distribution, reinforcing the notion that the majority of Olympic competitors belong to younger age brackets.

- The predominant age group identified fell within the 20 to 30-year-old bracket, **a range typically associated with the prime years of performance for athletes at their peak.**

Source:
https://sites.dartmouth.edu/sportsanalytics/2021/11/10/peak-age-in-sports/#:~:text=In%20general%2C%20research%20shows%20that,age%20of%2027%20or%2028.

Results Obtained

- The distribution shown in the **Height graphic** had values for **both Kurtosis and Skewness that were extremely close to zero.** This suggests that the distribution closely approximates a normal or bell-shaped curve.

- **Extreme outliers** in the dataset significantly affected measures like means, including extremely heavy/light, tall/small, and young/old athletes.

Source:
https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm#:~:text=Skewness%20is%20a%20measure%20of,relative%20to%20a%20normal%20distribution.

Future Works

Future Works

- A more in depth study looking at KPI’s for medal winners.
- Looking at performance of specific countries or regions in different disciplines.
- Looking for more association measures between different variables.
- Understanding the ideal metrics across different disciplines.

Link to Colab:
https://colab.research.google.com/drive/18ARJv1YMeyOPG2-0Y4FHHiAxS3eDzBR3?usp=sharing

Thank you