

|  |
| --- |
| Geograpical vs  Age\_SEX\_HEight\_Weight |
| Ronak Singh |
|  |
|  |

# Introduction

This report analyzes and segments Olympic athlete data based on geographic distribution, sex, age, height, and weight. The aim is to uncover demographic patterns, regional trends, and insights into physical attributes of athletes across different cities and countries.

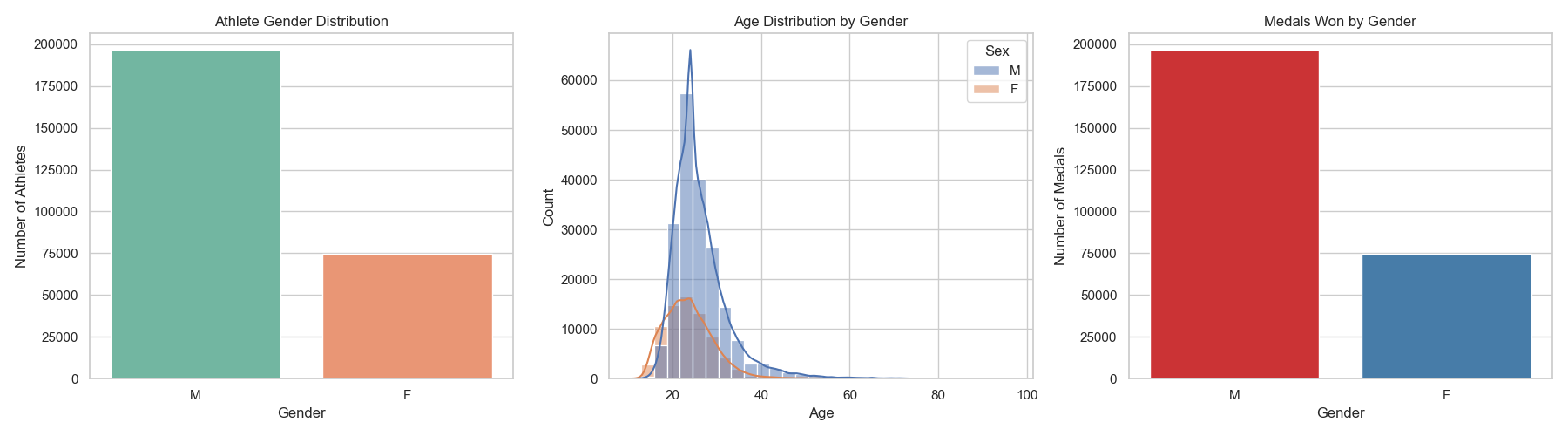
# Dataset overview

The dataset includes Olympic athlete records with features such as ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, and Medal. For this analysis, we focus on Sex, Age, Height, Weight, and geographic information (City, Team/NOC).

|  |  |
| --- | --- |
| Entries | 271116 |
| Winners | 39783 |
| Gold | 13372 |
| Silver | 13116 |
| Bronze | 13295 |

# Data Preparation

All records with missing values in the Sex, Age, Height, or Weight fields were removed to ensure data accuracy. Sex was encoded numerically (Male: 0, Female: 1) for clustering and analysis.



# Geograpic Segmentaion

Athletes were grouped by City and NOC (country). The geographic analysis revealed that cities with a higher athlete count tend to have a wider range of physical attributes, indicating diverse participation. NOC-based segmentation helped identify regional strengths and participation patterns in different sports disciplines.

# Demograpic segmentaion

Sex, Age, Height, and Weight were used to segment the athlete population. Descriptive statistics and visual exploration revealed common trends such as:

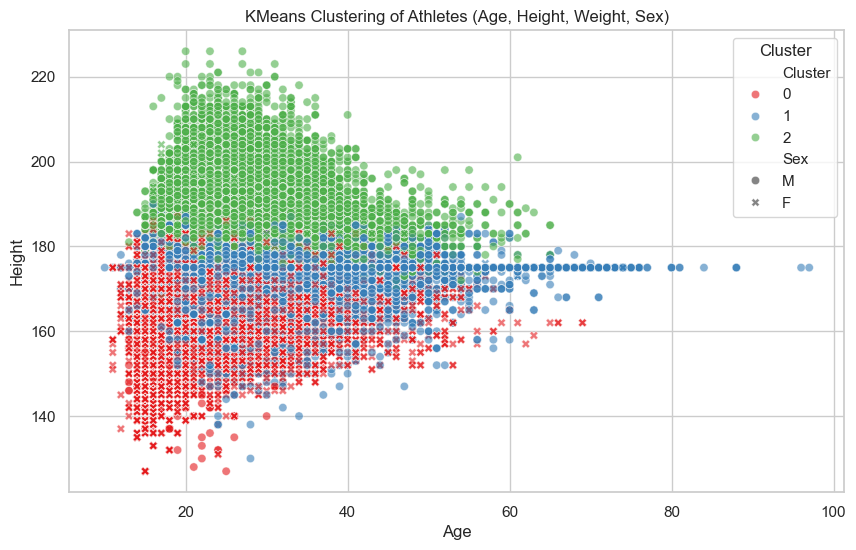
- Males generally had higher average height and weight than females.  
- Most athletes were between 20-30 years of age.  
- Weight and height distribution varied significantly by sport and region.

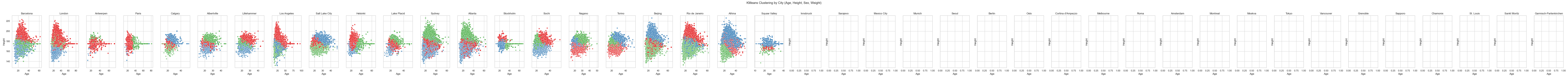


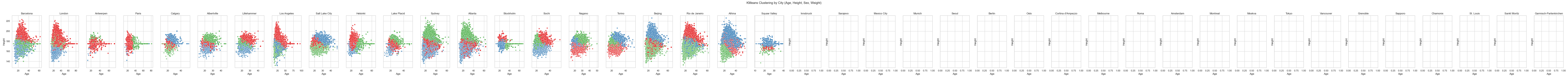
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | **F** | **M** | **Age** | **F** | **M** | **Age** | **F** | **M** | **Age** | **F** | | **M** |
| **10** | 0 | 1 | **30** | 2314 | 7174 | **50** | 30 | 248 | **70** | 10 | | 18 |
| **11** | 12 | 1 | **31** | 1780 | 5779 | **51** | 16 | 183 | **71** | 0 | | 33 |
| **12** | 32 | 7 | **32** | 1420 | 4826 | **52** | 29 | 215 | **72** | 3 | | 21 |
| **13** | 151 | 36 | **33** | 1064 | 3736 | **53** | 25 | 175 | **73** | 1 | | 7 |
| **14** | 744 | 93 | **34** | 856 | 3129 | **54** | 14 | 148 | **74** | 4 | | 8 |
| **15** | 1911 | 292 | **35** | 602 | 2531 | **55** | 10 | 94 | **75** | 0 | | 4 |
| **16** | 3024 | 828 | **36** | 469 | 2034 | **56** | 9 | 122 | **76** | 0 | | 7 |
| **17** | 3462 | 1914 | **37** | 345 | 1608 | **57** | 3 | 66 | **77** | 0 | | 2 |
| **18** | 4136 | 4016 | **38** | 262 | 1350 | **58** | 9 | 75 | **80** | 0 | | 3 |
| **19** | 4427 | 7216 | **39** | 186 | 1219 | **59** | 3 | 84 | **81** | 0 | | 2 |
| **20** | 4831 | 10427 | **40** | 156 | 1054 | **60** | 4 | 84 | **84** | 0 | | 1 |
| **21** | 5511 | 13653 | **41** | 125 | 828 | **61** | 4 | 64 | **88** | 0 | | 3 |
| **22** | 5416 | 15398 | **42** | 117 | 749 | **62** | 5 | 57 | **96** | 0 | | 1 |
| **23** | 5447 | 16428 | **43** | 91 | 702 | **63** | 10 | 46 | **97** | 0 | | 1 |
| **24** | 5715 | 25479 | **44** | 67 | 616 | **64** | 0 | 30 |
| **25** | 4950 | 14757 | **45** | 54 | 530 | **65** | 6 | 78 |  | |  | | |
| **26** | 4293 | 13382 | **46** | 57 | 372 | **66** | 3 | 28 |  | |  | | |
| **27** | 3971 | 12054 | **47** | 43 | 365 | **67** | 1 | 24 | **Sex** | | **Total\_Medals** | | |
| **28** | 3344 | 10699 | **48** | 31 | 376 | **68** | 7 | 18 | **F** | | 11253 | | |
| **29** | 2838 | 8625 | **49** | 46 | 316 | **69** | 16 | 44 | **M** | | 28530 | | |

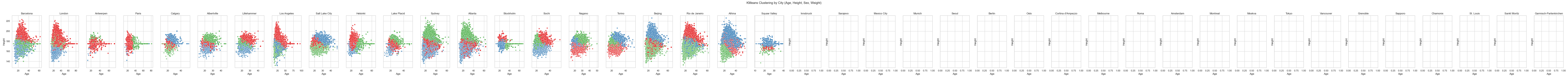
# Clustring Analysis

K-Means clustering was used to segment athletes based on Age, Height, Weight, and Sex. The model grouped athletes into distinct profiles, such as:  
- Cluster 1: Young, lightweight males  
- Cluster 2: Tall, heavyweight females  
- Cluster 3: Mid-aged athletes with average build  
This helped reveal common physical profiles within each city.









**Correlation Matrix:**

**Age Height Weight Sex**

**Age** 1.000000 0.098195 0.139615 -0.149095

**Height**  0.098195 1.000000 0.785382 -0.461667

**Weight**  0.139615 0.785382 1.000000 -0.476374

**Sex**  -0.149095 -0.461667 -0.476374 1.000000

# Key Insights

- Certain cities have consistent athlete body types, suggesting localized training norms.  
- Female participation is increasing over time and across more cities.  
- Height and weight distributions tend to cluster around specific age ranges.

**Most common city** London **Sex ratio** (0=Male, 1=Female): Sex

**Average Age** 25.81 0 0.686846

**Average Height** 175.61 cm 1 0.313154

**Average Weight** 70.82 kg

|  |  |  |  |
| --- | --- | --- | --- |
| **City** | **Participation\_Count** | **City** | **Participation\_Count** |
| *London* | 22426 | *Vancouver* | 4402 |
| *Athina* | 15556 | *Torino* | 4382 |
| *Sydney* | 13821 | *Stockholm* | 4338 |
| *Atlanta* | 13780 | *Antwerpen* | 4292 |
| *Rio de Janeirovvvvvvvvvvv* | 13688 | *Salt Lake City* | 4109 |
| *Beijing* | 13602 | *Innsbruck* | 3639 |
| *Barcelona* | 12977 | *Nagano* | 3605 |
| *Los Angeles* | 12423 | *Albertville* | 3436 |
| *Seoul* | 12037 | *Lillehammer* | 3160 |
| *Munich* | 10304 | *Calgary* | 2639 |
| *Montreal* | 8641 | *Sarajevo* | 2134 |
| *Mexico City* | 8588 | *Lake Placid* | 2098 |
| *Helsinki* | 8270 | *Grenoble* | 1891 |
| *Roma* | 8119 | *Sankt Moritz* | 1657 |
| *Tokyo* | 7702 | *Sapporo* | 1655 |
| *Moskva* | 7191 | *Cortina d'Ampezzo* | 1307 |
| *Paris* | 7169 | *St. Louis* | 1301 |
| *Berlin* | 6506 | *Squaw Valley* | 1116 |
| *Amsterdam* | 4992 | *Oslo* | 1088 |
| *Sochi* | 4891 | *Garmisch-Partenkirchen* | 895 |
| *Melbourne* | 4829 | *Chamonix* | 460 |