



UNIVERSITY OF WOLLONGONG AUSTRALIA

School of Mathematics & Applied Statistics

DSAA811: Data Analytics and Visualisation

Preliminary Report

Friday 28th March 2025 to Friday 11th April 2025

Sharon Van Den Berg 9251936 (stvdb914@uowmail.edu.au)

Friday 28th March 2025 to Friday 11th April 2025

DECLARATION

No part of this Assignment has been copied from anyone else, and I have not lent any part of it to anyone else. No part of this assignment has been written by generative AI.



Sharon Van Den Berg (9251936)

Tuesday 20th May, 2025

Date

This report was made using R Markdown as a project linking to my Git account

Abstract

- For now, just a heading for this section

Glossary

```
as_tibble(glossaryDef)
```

```
## # A tibble: 1 x 2
##   Acronym Meaning
##   <chr>    <chr>
## 1 NOC      National Olympic Committee
```

Contents

Abstract	2
Glossary	2
Introduction	4
Background	4
Research questions and aims of the project	4
Rationale	4
Data Description	5
Exploritory data analysis	6
Conclusion / Discussion	14
Bibliography	16

Introduction

- For now, just a heading for this section

Background

I have been looking into some articles regarding the statistics around reducing the number of athletes that a country can send to the Olympics to procure the maximum medals. The following three articles are of interest, and depending on the information, I can extract from my dataset. One of these avenues will become the focus element for this report.

(Haut, Prohl, and Emrich 2014) examines the statistics of the Olympics but also investigates the data from the perspective of increasing funds to rural areas to enhance performance and ultimately achieve more winners. At this stage I don't think that this report will be useful, however there is some regional information in my data set that I can use to plot the locations of the athletes on a world map.

(Condon, Golden, and Wasil 1999) Uses neural networks to produce three models that look at winners from a country perspective using data up until 1996. If I can get the same results, my data set continues to 2016, so this avenue would allow me to extend their findings on the years post 1996 to 2016.

(Heazlewood 2006) Looks into creating models to predict the optimal athlete numbers for all swimming events. This article was able to make some of these predictions, but improvements are needed to apply these results to athletics and swimming across the various distances of the races. These models are applied to results from 2004 and earlier. This article is of particular interest because of the break down of athletes into sections, mainly swimming. It could be useful to attempt to extend their results to apply it to other sports, such as athletics in summer or cross country skiing in winter.

Some of the terminology that I have come across is NOC or National Olympic Committee. Each country has a member that is on the committee and they are responsible for that country. Every athlete must compete under their NOC. I can use this NOC to get a general location of the athletes home town when comparing countries for athlete levels.

Research questions and aims of the project

As this is in the early stages of investigation, I still have too many questions to look into.

I can look at athletes, per winter or summer games, per country that the athlete competes at, or per NOC. I can even split the data into per sport. At the end of the day I want to answer, what is the minimal number of athletes that my country should send to the games to receive the most medals based on previous summer and winter Olympic data.

My plan moving forward is to play with the data set some more to see if I can marry the results of prior written research so I can stand on the giants of the past to push the research in this area forward in the future.

Rationale

It is no real stretch to underestimate the importance of pride that can come from winning many medals at an Olympic games. From the eyes of the country the cost to participate can be exorbitant to send one athlete, let alone an entire team of athletes. The rationale for this project is to maximize the number of medals that a country can

win, whilst reducing the costs of sending athletes to perform on this stage. I am looking for the optimal number of competing athletes to maximize the gold. In order to look into this problem, we can use past results in order to predict the future. I am unsure at this stage if we can look at this in the scope of the entire country or if we can reduce this to certain sporting events, such as swimming or track and field teams.

Data Description

The (Bansal 2021) data set called “Olympics_” was compiled by “Harsh Bansal” and was last updated 4 years ago. The dataset was uploaded and sourced from Kaggle (Keating et al. 2025). According to the site, there is only one owner with no DOI Citation, provenance or license. The restriction on the data is placed on it by Kaggle by way of citation of the owner “Harsh Bansal”. I am using this data at my own risk as it has not been authenticated or carefully curated.

The dataset contains 4 files, “athlete_events_data_dictionary.csv” contains 15 observations of 2 variables, “country_definitions.csv” contains 230 observations of 3 variables, “country_definitions_data_dictionary.csv” contains 3 observations of 2 variables, and “athlete_events.csv” containing 271,116 observations of 15 variables.

The “athlete_events.csv” file contains all athlete information of all the Olympic games dating from 1896 summer games and 1924 winter games up to and including the 2016 summer Olympic games. The following table outlines the variables contained within the set.

athletes

##	Field	Description
## 1	ID	Unique number for each athlete
## 2	Name	Athlete's name
## 3	Sex	Male (M) or Female (F)
## 4	Age	Integer
## 5	Height	In centimeters
## 6	Weight	In kilograms
## 7	Team	Team name
## 8	NOC	National Olympic Committee 3-letter code
## 9	Games	Year and season
## 10	Year	Integer
## 11	Season	Summer or Winter
## 12	City	Host city
## 13	Sport	Sport
## 14	Event	Event
## 15	Medal	Gold, Silver, Bronze, or NA

The variables that I am most interested in is the medal type, so as a country we can maximize receiving these. The country that the athlete is from so we can gain counts of participants in each prior games. This will allow us to work out the proportion of winners. The sport they participated in to break down the best results. Potentially the height and weight for some sports are equally important. This information will become clearer as further graphs and analysis is performed during the next 7 weeks.

In the athletes table there is a field called NOC which is the National Olympic City code that links to the country definitions that will allow for better groupings of data by country when linked to each other.

Exploritory data analysis

The first thing we should do with the datasets is to load them into r using the following code.

```
#Read in the 4 csv files
athletes <- read.csv('./data/athlete_events_data_dictionary.csv',
                     header = TRUE)
events <- read.csv('./data/athlete_events.csv',
                  header = TRUE)
countryDefdd<- read.csv('./data/country_definitions_data_dictionary.csv',
                       header = TRUE)
countryDef <- read.csv('./data/country_definitions.csv',
                      header = TRUE)

#Get a summary view of the data
summary(athletes)
```

```
##      Field      Description
## Length:15      Length:15
## Class :character Class :character
## Mode  :character Mode  :character
```

```
summary(events)
```

```
##      ID      Name      Sex      Age
## Min.   :      1 Length:271116 Length:271116 Min.   :10.00
## 1st Qu.:34643  Class :character Class :character 1st Qu.:21.00
## Median :68205  Mode  :character Mode  :character Median :24.00
## Mean   :68249                      Mean   :25.56
## 3rd Qu.:102097                      3rd Qu.:28.00
## Max.   :135571                      Max.   :97.00
##                                     NA's   :9474
##      Height      Weight      Team      NOC
## Min.   :127.0    Min.   : 25.0 Length:271116 Length:271116
## 1st Qu.:168.0    1st Qu.: 60.0 Class :character Class :character
## Median :175.0    Median : 70.0 Mode  :character Mode  :character
## Mean   :175.3    Mean   : 70.7
## 3rd Qu.:183.0    3rd Qu.: 79.0
## Max.   :226.0    Max.   :214.0
## NA's   :60171    NA's   :62875
##      Games      Year      Season      City
## Length:271116    Min.   :1896 Length:271116 Length:271116
## Class :character 1st Qu.:1960 Class :character Class :character
## Mode  :character Median :1988 Mode  :character Mode  :character
##                                     Mean   :1978
##                                     3rd Qu.:2002
##                                     Max.   :2016
##
```

```
##      Sport          Event          Medal
## Length:271116    Length:271116    Length:271116
## Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character
##
##
##
##
```

The athletes table is the meta data for the events table. There is a lot of missing data in the events table for height and weight of the athletes. NOC, sex, and year are categorical variables and have been coded as characters or numerical. These will need to be re coded into factors.

```
summary(countryDefdd)
```

```
##      Field          Description
## Length:3          Length:3
## Class :character Class :character
## Mode  :character Mode  :character
```

```
summary(countryDef)
```

```
##      NOC          region          notes
## Length:230      Length:230      Length:230
## Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character
```

CountryDefdd is the meta data for the countryDef file. These columns are NOC, the region that can be used for geospatial maps and the actual country name if the geospace location is unavailable. The countryDef is the data to represent this information.

Before I try to perform some explorations on the data it is imperative that we clean the data up a bit. Factoring the above variables will help with speed to process the data.

```
events$Sex <- factor(events$Sex,
                     levels = c("M","F"),
                     labels = c("M","F"))

#years <- distinct(events, Year)
#events$Year <- factor(events$Year,
#                      levels = years,
#                      labels = years)
#class(events$Year)
```

From here we can get a breakdown of the number of athletes that compete in each sport since the 2000 Summer games as shown below.

```
Summer <- events %>% filter (Season == "Summer") %>% filter (Year > 2000)
```

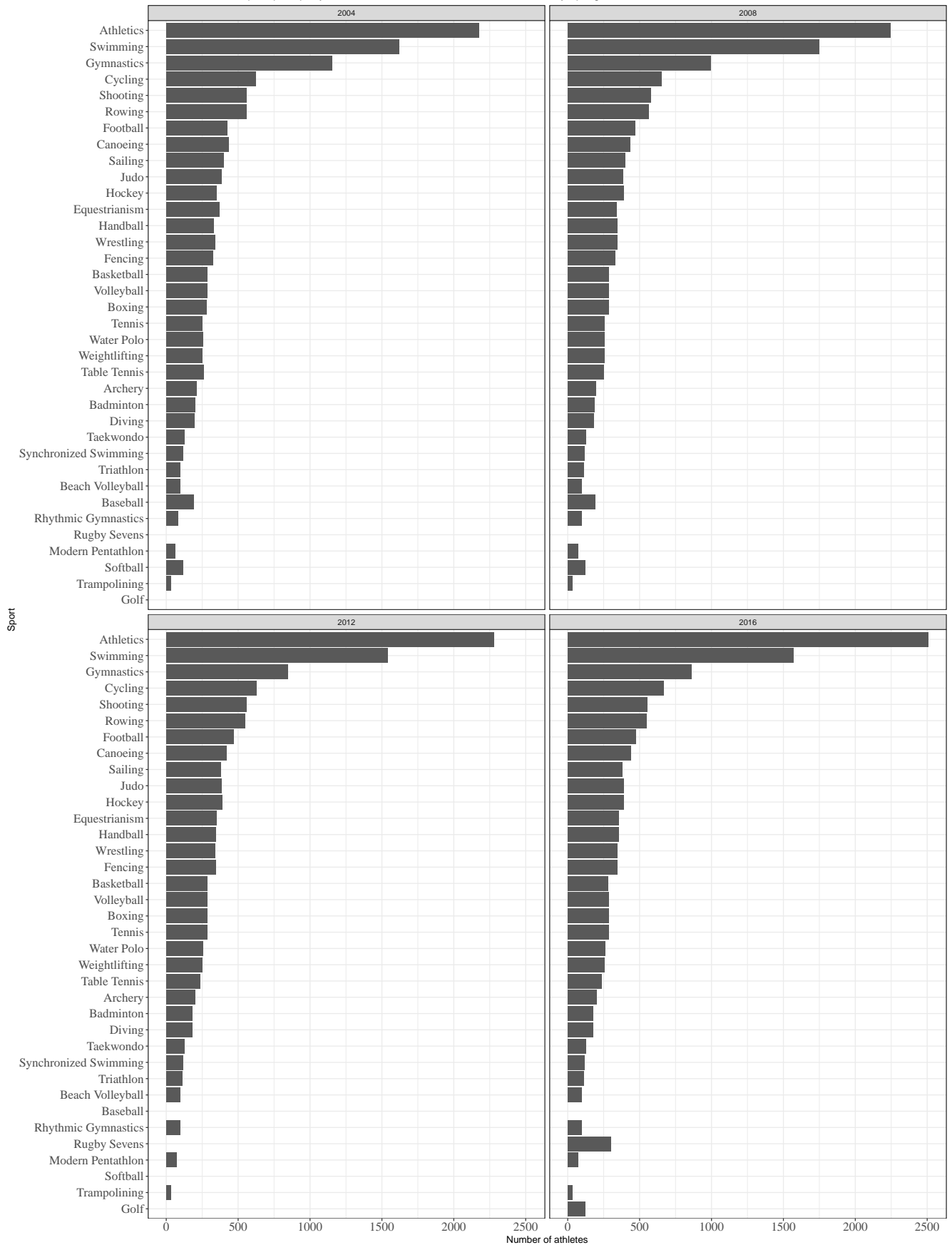
```
txtTitle <- paste('Number of athletes per sport, per year, between 2000 ',
                  'and 2020 at the summer olympic games')
```

```

Summer %>%
  ggplot() +
  geom_bar(aes(y = fct_rev(fct_infreq(Sport))), stat="count") +
  labs(title = txtTitle,
        x = "Number of athletes", y = "Sport") +
  theme_bw() +
  theme(axis.text = element_text(family="serif", size = 14)) +
  facet_wrap(vars(Summer$Year))

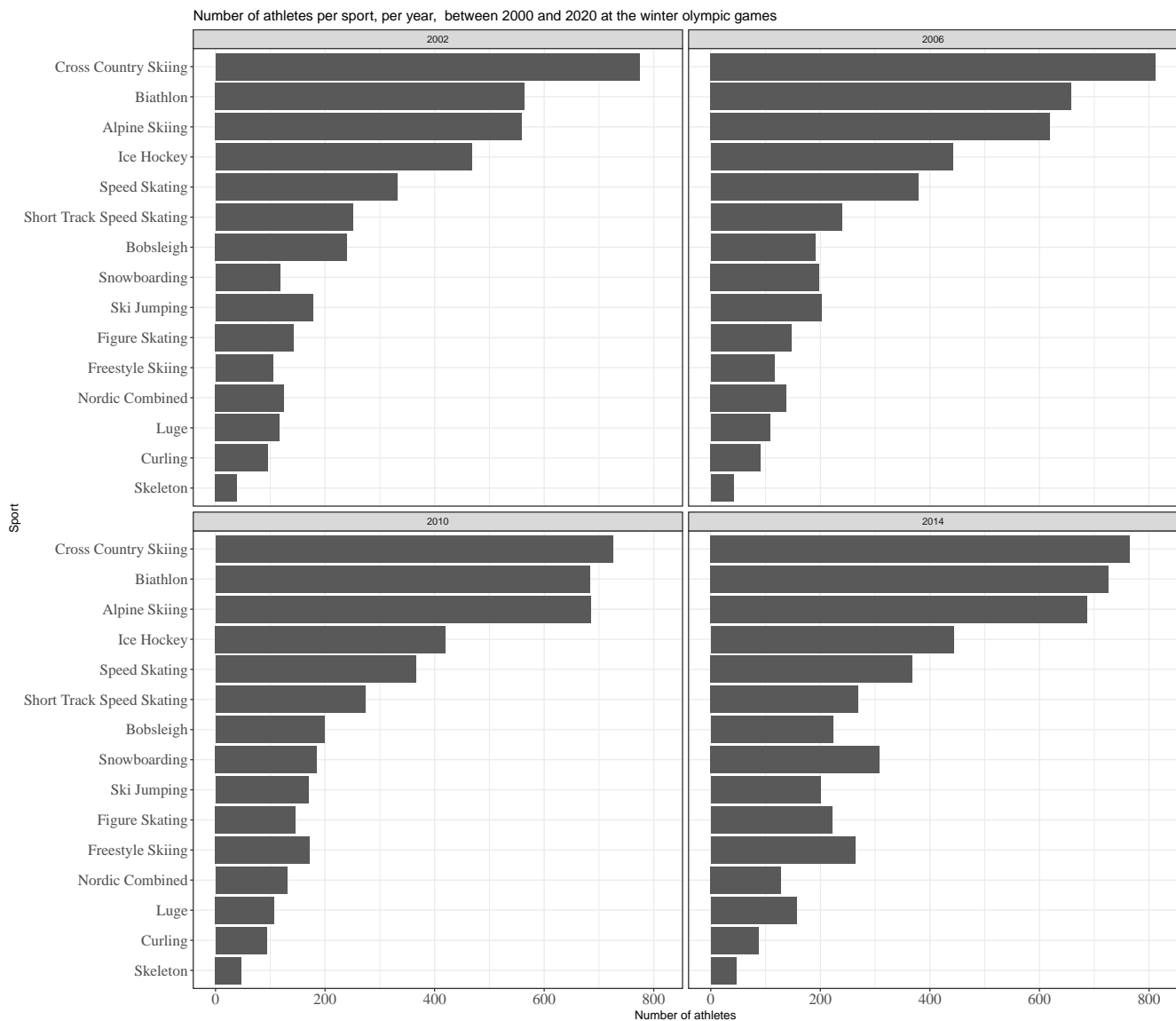
```


Number of athletes per sport, per year, between 2000 and 2020 at the summer olympic games



Similarly we can get a break down of the number of competing athletes at the Winter olympic games since 2000

```
txtTitle <- paste('Number of athletes per sport, per year,',
  ' between 2000 and 2020 at the winter olympic games')
winter %>%
  ggplot() +
  geom_bar(aes(y = fct_rev(fct_infreq(Sport))), stat="count") +
  labs(title = txtTitle,
    x = "Number of athletes", y = "Sport") +
  theme_bw() +
  theme(axis.text = element_text(family="serif", size = 14)) +
  facet_wrap(vars(winter$Year))
```



Whilst these plots look cool it is hard to read. It does take up a lot of space on the page, and I am not seeing any major changes to the number of athletes that have been sent in the past four summer or winter olympics. If this data is needed a table would be a better representation of it. Below is the same information for one of the olympics, summer or winter.

```
summerCounts <- Summer %>%
  group_by (Year,Sport) %>%
  count(Sport) %>%
  arrange(Year, desc(n)) %>%
  pivot_wider(values_from = n, names_from = Sport)

summerCounts <- rotate_df(summerCounts)
print.data.frame (summerCounts)
```

##	V1	V2	V3	V4
## Year	2004	2008	2012	2016
## Athletics	2175	2244	2278	2508
## Swimming	1618	1749	1538	1568
## Gymnastics	1151	996	848	861
## Cycling	623	652	629	667
## Rowing	559	563	550	550
## Shooting	558	577	560	555
## Canoeing	434	436	418	441
## Football	425	469	467	473
## Sailing	401	400	379	380
## Judo	384	386	384	389
## Equestrianism	371	340	350	355
## Hockey	352	387	387	390
## Wrestling	342	343	339	346
## Handball	328	343	347	353
## Fencing	323	332	345	346
## Basketball	287	287	287	281
## Volleyball	283	283	287	283
## Boxing	280	283	283	283
## Table Tennis	260	250	236	236
## Water Polo	255	256	257	258
## Tennis	252	254	286	286
## Weightlifting	249	253	252	255
## Archery	212	194	200	200
## Badminton	200	184	182	177
## Diving	196	182	181	178
## Baseball	191	191	NA	NA
## Taekwondo	124	126	128	126
## Softball	118	120	NA	NA
## Synchronized Swimming	117	117	117	118
## Triathlon	99	110	110	110
## Beach Volleyball	96	96	96	96
## Rhythmic Gymnastics	84	95	95	96
## Modern Pentathlon	64	72	72	72
## Trampolineing	32	32	32	32

```
## Rugby Sevens      NA    NA    NA  299
## Golf               NA    NA    NA  120
```

```
winterCounts <- winter %>%
  group_by (Year, Sport) %>%
  count(Sport) %>%
  arrange(Year, desc(n)) %>%
  pivot_wider(values_from = n, names_from = Sport)

winterCounts <- rotate_df(winterCounts)
print.data.frame (winterCounts)
```

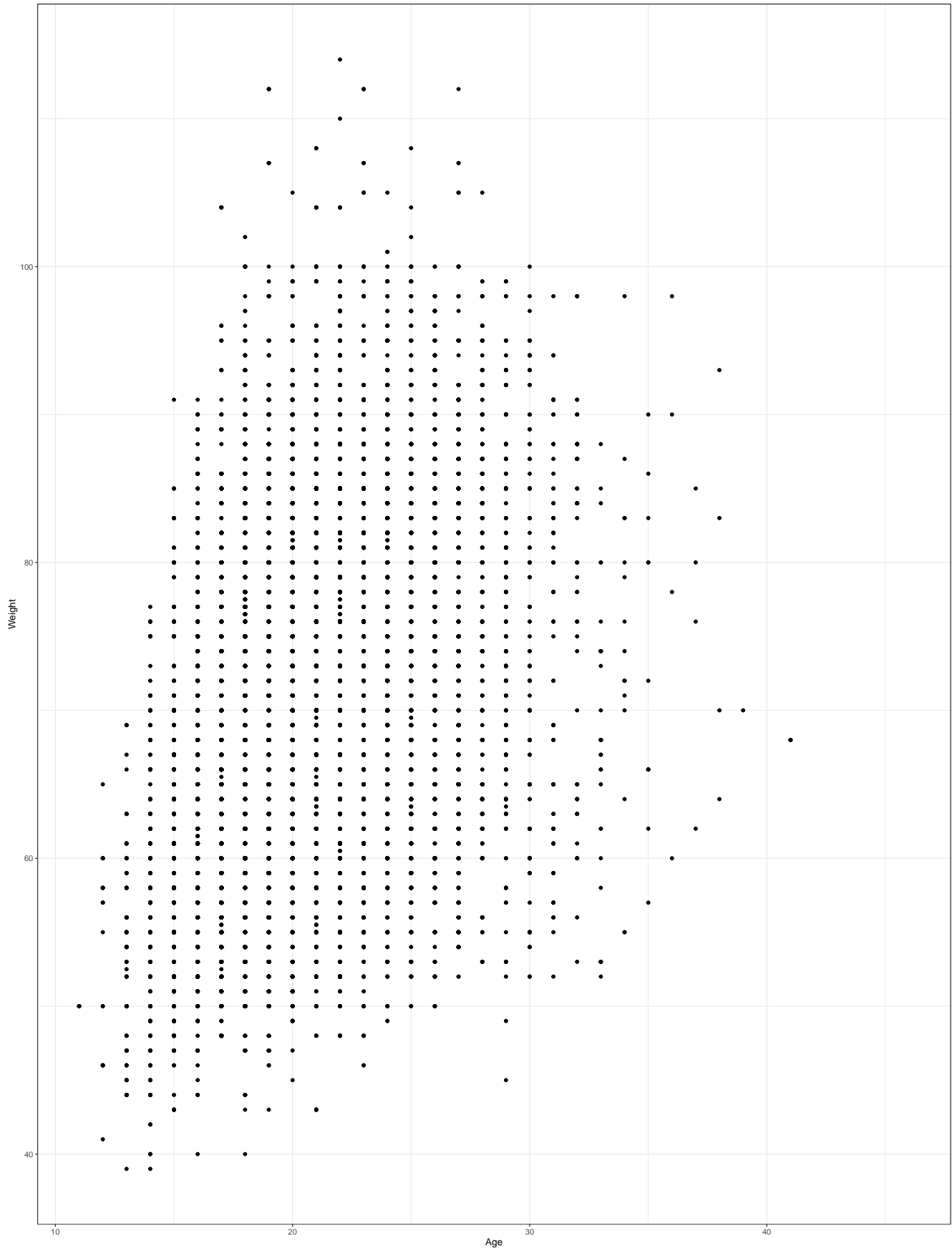
```
##           V1    V2    V3    V4
## Year      2002 2006 2010 2014
## Cross Country Skiing  774  812  725  765
## Biathlon           564  658  683  726
## Alpine Skiing       559  619  685  687
## Ice Hockey          468  442  420  443
## Speed Skating       332  379  366  368
## Short Track Speed Skating 251  239  273  269
## Bobsleigh           240  191  199  223
## Ski Jumping         178  202  170  200
## Figure Skating      143  147  146  221
## Nordic Combined     125  138  131  128
## Snowboarding        118  198  185  308
## Luge                117  108  107  156
## Freestyle Skiing    105  116  172  264
## Curling              96   91   93   87
## Skeleton            39   42   47   46
```

This data set lets you plot the athletes height and weight to see if there is any benefit to selecting athletes of a certain structure. Below is an example using the swimming data in the data set

```
events %>% filter(Sport == "Swimming") %>%
  ggplot() +
  geom_point(aes(x = Age, y = Weight)) +
  labs(title = 'Olympic swimmers ages and weights') +
  theme_bw()
```

```
## Warning: Removed 4411 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

Olympic swimmers ages and weights



From this we can see that swimmers are optimal when they are under the age of 40 and smaller than 100 kilograms in general, across all the Olympics.

At this stage I still need to find some more research articles that I can replicate to make a story with this data set, There are still many avenues that I can follow up with.

Conclusion / Discussion

- For now, just a heading for this section

#Session Information

sessionInfo()

```
## R version 4.3.2 (2023-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 26100)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: Australia/Sydney
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] sjmisc_2.8.10   lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1
## [5] dplyr_1.1.4     purrr_1.0.4     readr_2.1.5     tidyr_1.3.1
## [9] tibble_3.2.1    ggplot2_3.5.2   tidyverse_2.0.0 tinytex_0.57
## [13] knitr_1.50
##
## loaded via a namespace (and not attached):
## [1] gtable_0.3.6      compiler_4.3.2    tidyselect_1.2.1  dichromat_2.0-0.1
## [5] scales_1.4.0      yaml_2.3.10       fastmap_1.2.0     R6_2.6.1
## [9] labeling_0.4.3    generics_0.1.4    sjlabelled_1.2.0  insight_1.2.0
## [13] rprojroot_2.0.4   pillar_1.10.2     RColorBrewer_1.1-3 tzdb_0.5.0
## [17] rlang_1.1.5       utf8_1.2.4        stringi_1.8.7     xfun_0.51
## [21] timechange_0.3.0  cli_3.6.2         withr_3.0.2       magrittr_2.0.3
## [25] digest_0.6.37     grid_4.3.2        rstudioapi_0.17.1 hms_1.1.3
## [29] lifecycle_1.0.4   vctrs_0.6.5       evaluate_1.0.3    glue_1.8.0
## [33] farver_2.1.2      rmarkdown_2.29    tools_4.3.2       pkgconfig_2.0.3
## [37] htmltools_0.5.8.1
```

Bibliography

- Bansal, Harsh. 2021. “Olympics_.” Kaggle.com. <https://www.kaggle.com/datasets/harshbansal27/olympics>.
- Condon, Edward M, Bruce L Golden, and Edward A Wasil. 1999. “Predicting the Success of Nations at the Summer Olympics Using Neural Networks.” *Computers & Operations Research* 26 (November): 1243–65. [https://doi.org/10.1016/s0305-0548\(99\)00003-9](https://doi.org/10.1016/s0305-0548(99)00003-9).
- Haut, Jan, Robert Prohl, and Eike Emrich. 2014. “Nothing but Medals? Attitudes Towards the Importance of Olympic Success.” *International Review for the Sociology of Sport* 51 (March): 332–48. <https://doi.org/10.1177/1012690214526400>.
- Heazlewood, Timothy. 2006. “Prediction Versus Reality: The Use of Mathematical Models to Predict Elite Performance in Swimming and Athletics at the Olympic Games.” *Journal of Sports Science & Medicine* 5 (December): 480. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3861753/>.
- Keating, Nate, Jeff Moser, William Cukierski, Jerad Rose, Myles O’Neill, Risdal Meg, Meghan O’Connell, et al. 2025. “Kaggle: Your Home for Data Science.” Kaggle.com. <https://www.kaggle.com/>.