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)

Thursday 10th April, 2025

Date

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

Introduction

- For now, just a heading for this section

Background

This is the sub section of the main project

(Haut, Prohl, and Emrich 2014) Looks into the statistics of the Olympics but is investigating the data from the perspective of increasing funds to the rural areas to increase performance and then winners. (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 (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.

Research questions and aims of the project

This is the research questions part

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 rational 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.

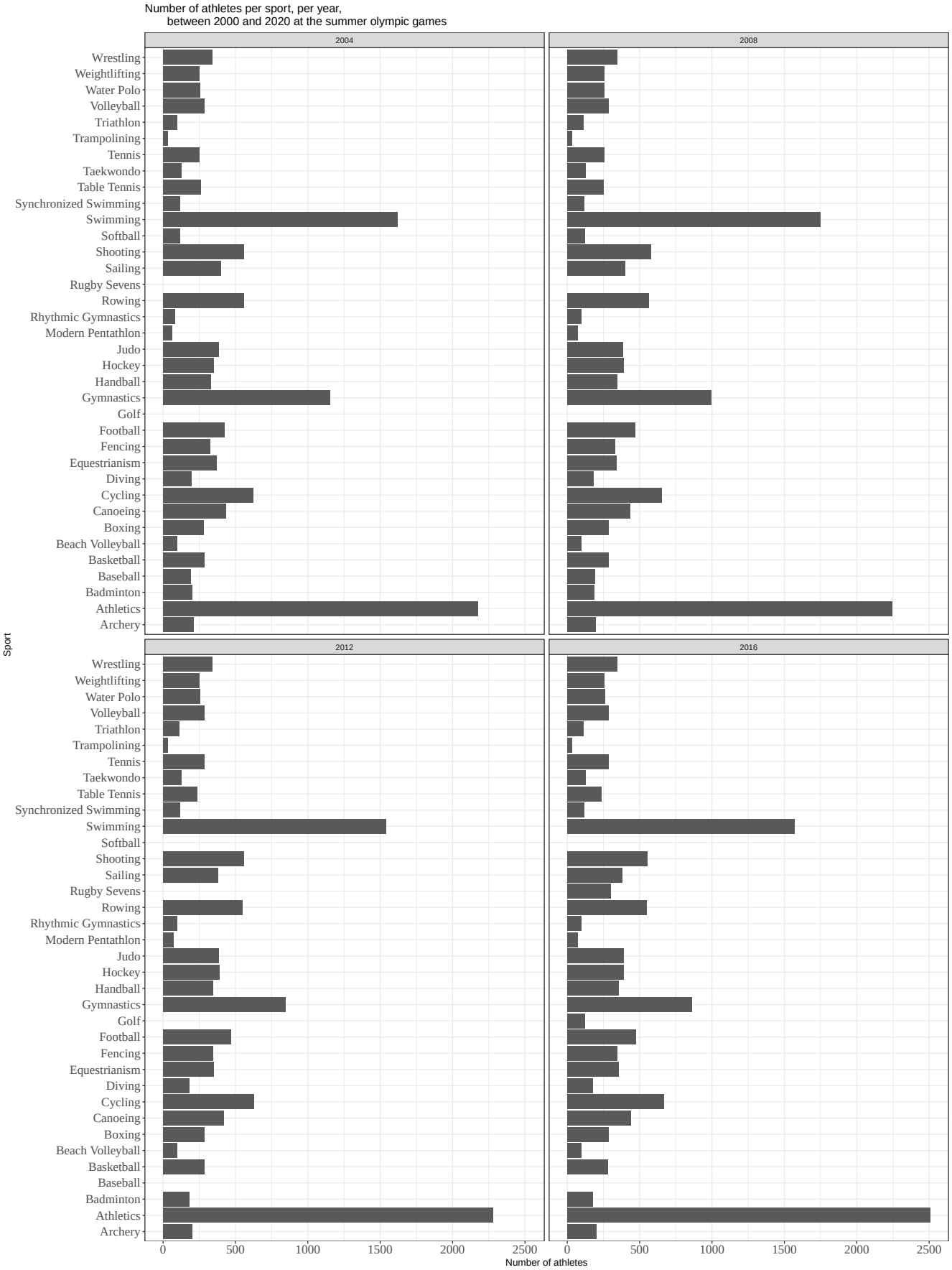
```
events$Sex <- factor(events$Sex,
                     levels = c("M","F"),
                     labels = c("M","F"))

#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)
```

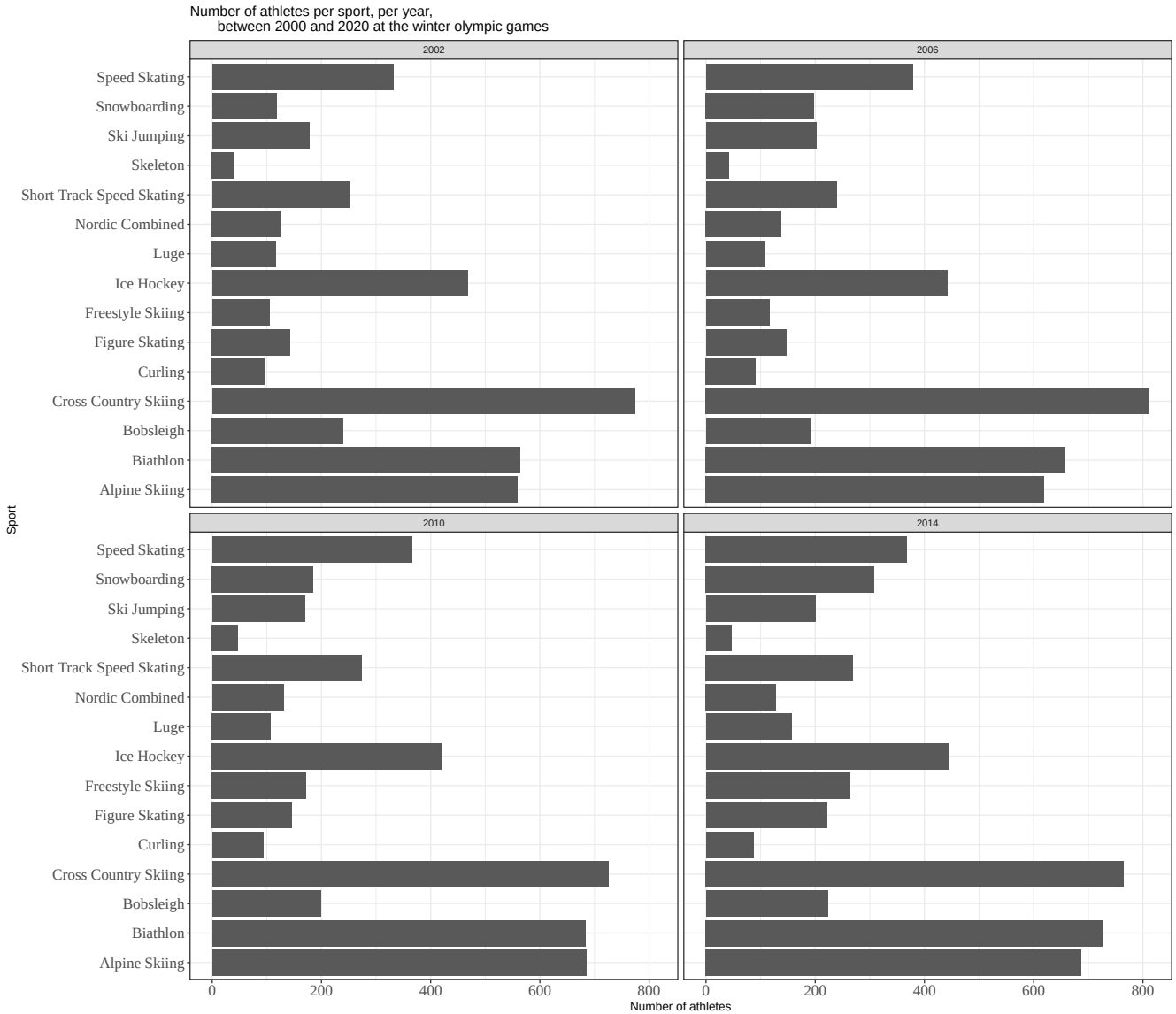
```
Summer %>%
  ggplot() +
  geom_bar(aes(y = Sport), stat="count") +
  labs(title = 'Number of athletes per sport, per year,
       between 2000 and 2020 at the summer olympic games',
       x = "Number of athletes", y = "Sport") +
  theme_bw() +
  theme(axis.text = element_text(family="serif", size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


Number of athletes per sport, per year,
between 2000 and 2020 at the summer olympic games
2004
2008
2012
2016
Sport
Wrestling
Weightlifting
Water Polo
Volleyball
Triathlon
Trampolining
Tennis
Taekwondo
Table Tennis
Synchronized Swimming
Swimming
Softball
Shooting
Sailing
Rugby Sevens
Rowing
Rhythmic Gymnastics
Modern Pentathlon
Judo
Hockey
Handball
Gymnastics
Golf
Football
Fencing
Equestrianism
Diving
Cycling
Canoeing
Boxing
Beach Volleyball
Basketball
Baseball
Badminton
Athletics
Archery
0
500
1000
1500
2000
2500
Number of athletes

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

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

```
## Warning in geom_bar(aes(y = Sport, stat = "count")): Ignoring unknown
## aesthetics: stat
```



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_Australia.utf8  LC_CTYPE=English_Australia.utf8
## [3] LC_MONETARY=English_Australia.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_Australia.utf8
##
## time zone: Australia/Sydney
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.4     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.1   tidyverse_2.0.0 tinytex_0.56    knitr_1.50
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.6      compiler_4.3.2    tidyselect_1.2.1  scales_1.3.0
##  [5] yaml_2.3.10       fastmap_1.2.0     R6_2.6.1          labeling_0.4.3
##  [9] generics_0.1.3    munsell_0.5.1     rprojroot_2.0.4   pillar_1.10.1
## [13] tzdb_0.5.0        rlang_1.1.5       utf8_1.2.4        stringi_1.8.7
## [17] xfun_0.51         timechange_0.3.0  cli_3.6.2         withr_3.0.2
## [21] magrittr_2.0.3    digest_0.6.37     grid_4.3.2        rstudioapi_0.17.1
## [25] hms_1.1.3         lifecycle_1.0.4   vctrs_0.6.5       evaluate_1.0.3
## [29] glue_1.8.0        farver_2.1.2      colorspace_2.1-1  rmarkdown_2.29
## [33] tools_4.3.2       pkgconfig_2.0.3   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.

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/.