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 athletic and swimming across the various distances of the races.

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
```

```
##      ID           Unique number for each athlete
##    Name                            Athlete's name
##     Sex                  Male (M) or Female (F)
##     Age                                 Integer
##  Height                           In centimeters
##  Weight                             In kilograms
##    Team                                Team name
##     NOC National Olympic Committee 3-letter code
##   Games                           Year and season
##    Year                                 Integer
## Season                          Summer or Winter
##    City                                Host city
##   Sport                                    Sport
##   Event                                    Event
##   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

Exploritory # 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          generics_0.1.3
##  [9] munsell_0.5.1     rprojroot_2.0.4   pillar_1.10.1     tzdb_0.5.0
## [13] rlang_1.1.5       utf8_1.2.4        stringi_1.8.7     xfun_0.51
## [17] timechange_0.3.0  cli_3.6.2         withr_3.0.2       magrittr_2.0.3
## [21] digest_0.6.37     grid_4.3.2        rstudioapi_0.17.1 hms_1.1.3
## [25] lifecycle_1.0.4   vctrs_0.6.5       evaluate_1.0.3    glue_1.8.0
## [29] colorspace_2.1-1  rmarkdown_2.29    tools_4.3.2       pkgconfig_2.0.3
## [33] 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/.