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)

Wednesday 9th April, 2025

Date

Abstract

- For now, just a heading for this section

Contents

Introduction

- For now, just a heading for this section

Background

This is the sub section of the main project glossary_popup("hover") glossary("NOC") glossary("p-value", show = "def")

glossary_table()

Research questions and aims of the project

This is the research questions part

Rationale

This is the rational

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

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] glossary_1.0.0  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_3.5.1   tidyverse_2.0.0 tinytex_0.56
## [13] 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       stringi_1.8.7     xfun_0.51         timechange_0.3.0
## [17] cli_3.6.2         withr_3.0.2       magrittr_2.0.3    digest_0.6.37
## [21] grid_4.3.2        rstudioapi_0.17.1 hms_1.1.3         lifecycle_1.0.4
## [25] vctrs_0.6.5       evaluate_1.0.3    glue_1.8.0        colorspace_2.1-1
## [29] rmarkdown_2.29    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.

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/.