R. STRAIGHT

MIS 545 PROJECT

Introduction

The Olympics has become a financial juggernaut from the viewpoint of both the hosting country and the value of being a medalist. In 2016, the Rio Olymipcs cost $13.1 billion US dollars (7.23 billion reals) to host.[1]. This included “a subway line, a doping laboratory, a renovated port and cleanup of pouted Guanabara Bay.”

The value of *gold* is more than the value of the medal, of course. Countries reward their medalists depending on which medal they bring home and these rewards vary drastically from country to country. Singapore, for example, rewards gold medalists with $1 million USD while Canada pays a comparatively paltry $15,000 USD[2]. Advertising sales during the Rio Olympics in 2016 totaled $1.2 billion USD.[3]

[1] Rio Olympics cost $13.1 billion

[2] Here’s how much Olympic athletes earn in 12 different countries

[3] NBC says it has topped $1 billion in national ad sales for 2020 Summer Olympics

Needless to say, it behooves interested parties to be able to predict just who, when, and where medalists will crop up, whether this is in an effort to determine if the 12-year-old male gymnast in Sweden is likely to be a 16-year-old gold medalist in four years or if this or that country is worth scouting in for talent given their past medal winnings.

The Chosen Data

The purpose of this project is to apply a variety of classification and predictive methodologies to a chosen data set for the purposes of demonstrating knowledge and skills developed throughout the semester. The dataset chosen for this project is 120 years of Olympic history: athletes and results[4] This particular dataset was chosen for a variety of reasons:

[4] Described by the creator as "basic bio data on athletes and medal results from Athens 1896 to Rio 2016."

- It is relatively large, coming in 271,116 rows when loaded raw.
- There is a variety of variable types to work with, providing a range of options when it comes to different classification and preditive tests.
- It affords a certain level of approachability and familiarity by virtue of its content; after all, we all know Olympic medalists.

The purpose of this study, then, is to examine the particulars of the Olympic historical record and attempt to identify trends and make predictions thereby. Three possibilities for this data in this context come to mind:

1. What trends are apparent in nations' medal totals?
2. What demographics contribute to medaling?
3. Can we predict a medal based on a collection of an athlete's demographics?

A Description of the Data

The data originates in a Kaggle.com dataset provided by Randi Griffin. According to Griffin,

> This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016 [scraped from] www.sports-reference.com in May 2018…. Note that the Winter and Summer Games were held in the same year up until 1992. After that, they staggered them such that Winter Games occur on a four year cycle starting with 1994, then Summer in 1996, then Winter in 1998, and so on. A common mistake people make when analyzing this data

is to assume that the Summer and Winter Games have always been staggered.

— Randi Griffin

The dataset is delivered in two files: `athlete_events.csv` and `noc_regions.csv`. The descriptions are provided in the data source.

File	Variable	Data type	Data format	Description
athlete_events.csv				
	ID	ind	int	Unique number for each athlete
	Name	ind	chr	Athlete's name
	Sex	dep	chr	M or F
	Age	ind	int	Integer
	Height	ind	int	In centimeters
	Weight	ind	num	In kilograms
	Team	ind	chr	Team name
	NOC	ind	chr	National Olympic Committee 3 letter code
	Games	ind	chr	Year and season
	Year	ind	int	Integer
	Season	ind	chr	Summer or Winter
	City	ind	chr	City
	Sport	ind	chr	Sport
	Event	ind	chr	Event
	Medal	dep	chr	Gold, Silver, Bronze, or `NA`
noc_regions.csv				
	NOC	ind	chr	National Olympic Committee 3 letter code
	region	ind	chr	Country name (matches with regions in `map_data("world")`)
	notes	ind	chr	Notes

Data Pre-Processing

Fortunately, Griffin's scraping techniques prove tidy and in need of very little cleaning, all things considered. The entirety of the loading and tidying is as follows:

```
# tidy up the titles
athlete_events <- athlete_events %>% clean_names()
noc_regions <- noc_regions %>% clean_names()

# Join up athlete_events and noc_regions to get a nice country name
olympics <- as_tibble(athlete_events %>% left_join(noc_regions, by = "noc"))

# Switch to factors
olympics <- olympics %>%
  mutate(across(c("sex", "team", "noc", "games", "year", "sport", "medal", "city",
    "region", "season"), factor))

# Replace NAs in "medal" with "None"
olympics$medal <- olympics$medal %>%
  as.character() %>%
  replace_na("none") %>%
  as_factor()

# There are way too many sports and a few only happened a couple times.
# Pare those down to the top 50, naming the rest "Other."
olympics$sport <- olympics$sport %>%
  fct_lump_n(n = 51)
```

With this we can now explore the data a bit easier. Note that each row in this dataset is for an *athlete*. Follows is a glimpse at the structure of the numerical data to verify data formats are as expected.

Table 2: Summary table of relevant data

age	height	weight	year	medal
Min. :10.00	Min. :127.0	Min. : 25.0	1992 : 16413	none :231333
1st Qu.:21.00	1st Qu.:168.0	1st Qu.: 60.0	1988 : 14676	Gold : 13372
Median :24.00	Median :175.0	Median : 70.0	2000 : 13821	Bronze: 13295
Mean :25.56	Mean :175.3	Mean : 70.7	1996 : 13780	Silver: 13116
3rd Qu.:28.00	3rd Qu.:183.0	3rd Qu.: 79.0	2016 : 13688	NA
Max. :97.00	Max. :226.0	Max. :214.0	2008 : 13602	NA
NA's :9474	NA's :60171	NA's :62875	(Other):185136	NA

Descriptive Analysis