Data Appendix

Kiersten Hamby,Breanna Ranglall,Krutika Tekwani

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.1     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(skimr)
```

Data Appendix

Group Members

Kiersten Hamby, Breanna Ranglall, Krutika Tekwani

Loading Data In

```
Athletics=read.csv("~/OneDrive - University of St. Thomas/STAT320/GroupE2AthleticsData/Gro
```

Skim

```
Athletics%>%
  select(institution_name,state_cd,classification_name,EFMaleCount,EFFemaleCount,EFTotalCo
  skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	2027
Number of columns	17
_____	
Column type frequency:	
character	4
numeric	13
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
institution_name	0	1	1	65	0	2009	0
state_cd	0	1	2	2	0	53	0
classification_name	0	1	4	34	0	19	0
sector_name	0	1	14	35	0	5	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
EFMaleCount	0	1.00	1747.77	2848.67	1	410.50	776.0	1720.00	34767	
EFFemaleCount	0	1.00	2186.75	3255.42	4	510.00	1022.0	2308.00	29793	
EFTotalCount	0	1.00	3934.52	6039.83	62	958.50	1796.0	4024.00	64560	
STUDENTAID_MEN	13	0.99	1239958.42	2131097.77	7	0.00	225186.5	1584597.00	15953475	
STUDENTAID_WOMEN	30	0.98	1046035.47	1796566.78	8	0.00	213148.0	1315407.75	14979930	
STUDENTAID_TOTAL	0	1.00	2271812.66	3871922.20	20	0.00	433850.0	3053948.00	28909727	
HDCOACH_SALARY_MEN	14	0.99	102533.65	256976.87	0	20000.00	39542.0	63980.00	2387189	
HDCOACH_SALARY_WOMEN	32	0.98	47082.84	52978.37	0	19267.00	34426.0	53592.50	601582	
RECRUITEXP_MEN	13	0.99	114303.40	379461.79	0	1075.00	13772.5	52260.25	5331795	
RECRUITEXP_WOMEN	30	0.98	46547.08	110694.45	0	1235.75	10646.0	31645.25	962902	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
RECRUITEXP_TOTAL	0	1.00	159561.80	481956.40	0	2412.50	25578.0	82591.50	6294697	
NUM_HDCOACH_MEN	15	0.99	6.28	3.77	1	4.00	6.0	8.00	106	
NUM_HDCOACH_WOMEN	33	0.98	7.11	3.91	0	5.00	7.0	9.00	101	

Variable Names/Types

institution_name

state_cd

classification_name

EFMaleCount

EFFemaleCount

EFTotalCount

sector_name

STUDENTAID_MEN: 13 missing values

STUDENTAID_WOMEN: 31 missing values

STUDENTAID_TOTAL

HDCOACH_SALARY_MEN: 14 missing values

HDCOACH_SALARY_WOMEN: 32 missing values

RECRUITEXP_MEN: 13 missing values

RECRUITEXP_WOMEN: 31 missing values

RECRUITEXP_TOTAL

NUM_HDCOACH_MEN: 15 missing values

NUM_HDCOACH_WOMEN: 33 missing values

The variable names are readable and clear. One point of confusion is what 'EF' means in the count of students. Sector name refers to the public/private status of a school. It is convenient that the data already specifies men's, women's, and totals for the variables. Some variables include underscores in the name.

All variables except those discussed below, Categorical, are numeric variables which is what they should be.

Category Names

institution_name

state_cd

classification_name

sector_name

All non-binary categorical variables. Clearly named, though they do include underscores. These are all classified as character variables which is what we would expect.

Max and Min for Variables

EFMaleCount: Max=34,767, Min=1

EFFemaleCount: Max=29,793, Min=4

EFTotalCount: Max=64,560, Min=62

STUDENTAID_MEN: Max=15,953,475, Min=0

STUDENTAID_WOMEN: Max=14,979,930, Min=0

STUDENTAID_TOTAL: Max=28,909,727, Min=0

HDCOACH_SALARY_MEN: Max=2,387,189, Min=0

HDCOACH_SALARY_WOMEN: Max=601,582, Min=0

RECRUITEXP_MEN: Max=5,331,795, Min=0

RECRUITEXP_WOMEN: Max=962,902, Min=0

RECRUITEXP_TOTAL: Max=6,294,697, Min=0

NUM_HDCOACH_MEN: Max=106, Min=1

NUM_HDCOACH_WOMEN: Max=101, Min=2

The only point of concern is that so many variables have a minimum of 0 and we would not expect to see a university with no recruiting expenses, for example. We will look into what is affecting this, potentially the sector of the school.

Data Wrangling, Data Cleaning

-Only have the parent school if there are multiple campuses to maintain independence

-Deleting unused variables as there are more than 4,000 variables

-Look into schools that have 0 as a minimum for expenses/student aid/etc. to see if we can find a cause or need to remove these observations from our data.