Final Project Assignment

Weizi Wu

2023-11-23

1.Create a R Markdown

As displayed, the title is "Final Project Assignment," authored by myself, with the date of when I started this assignment. The output format is PDF.

2. Course Review

Throughout this class, I have acquired valuable skills in using R for data manipulation, visualization and communication. The knowledge I have gained has equipped me with the capability to conduct data analysis using R and visualization techniques, which marks a new beginning for my future data analysis endeavors.

3. Setting up working directory

Before any data analysis, it's crucial to set the working directory to tell R where to look for files. Here, Knitr::_chunk$set() is used to specify any global settings to be applied to the R Markdown script.

```
knitr::opts_chunk$set(include = TRUE)
setwd("/Users/weiziwu/Desktop/Reproducibility_in_R/Chapter_5_HW")
```

4.Packages installing or calling up

Next, we installed or called up the necessary packages. It's important to note that we only need to install a package once, but we need to call it every time we reopen R. If you're using a package for the first time, please remove the "#" annotation and run the installation code before loading the package for usage. For this project, I used four packages: tidyverse (for data tidying), Psycho (for descriptive analysis), ggplot2 (for plotting), and readxl (for importing data). Therefore, the command library() was used to tell R these packages I planed to use and then they were loaded into my library.

```
# install.packages("tidyr")
# install.packages("psych")
# install.packages("ggplot2")
# install.packages("readr")
library(readxl)
library(ggplot2)
library(tidyverse)
library(psych)
```

5. Importing data and rename them

Here, since the data is in xlsx format, we can use the read_xlsx function to import it. If the file is in a different format, such as csv or xls, we can use the read_csv or read_xls function instead. Once imported, we use the name assignment -> to rename the data for analysis.

```
read_xlsx("/Users/weiziwu/Desktop/EPSY-5195/dataset/Single-sex Athletics.xlsx")->adata
read_xlsx("/Users/weiziwu/Desktop/EPSY-5195/dataset/Single-sex Classes.xlsx")->cdata
```

6. Tidy up

Here, the problem of the variable "MIN-MAX" is Two variables in one column, so I used the function separate(data, "var1_var2", into=c(var1, var2)) to split the variable MIN and MAX into two columns and assign new names. I used colnames() to check the result.

```
separate(adata, "MIN-MAX", into = c("MIN", "MAX"))->adata1
colnames(adata1)
```

```
 [1] "LEA_STATE"           "LEA_STATE_NAME"      "LEAID"
 [4] "LEA_NAME"            "SCHID"               "SCH_NAME"
 [7] "COMBOKEY"            "JJ"                  "SCH_SSATHLETICS_IND"
[10] "SCH_SSSPORTS_M"      "SCH_SSSPORTS_F"      "TOT_SSSPORTS"
[13] "SCH_SSTEAMS_M"       "SCH_SSTEAMS_F"       "TOT_SSTEAMS"
[16] "SCH_SSPART_M"        "SCH_SSPART_F"        "TOT_SSPART"
[19] "MIN"                 "MAX"
```

7. Creating new variable

Here, I planned to create a new variable to count the total number of single-sex classes. To achieve this, I used the function mutate(data, new_var = var1 + var2 +...) to sum up the totals of the existing columns. Next, I verified the new variable by using function colnames().

```
mutate(cdata, TOT_SSCLASSES=TOT_SSCLASSES_ALGG + TOT_SSCLASSES_OTHM + TOT_SSCLASSES_SCI +
colnames(cdata1)
```

```
 [1] "LEA_STATE"            "LEA_STATE_NAME"       "LEAID"
 [4] "LEA_NAME"             "SCHID"                "SCH_NAME"
 [7] "COMBOKEY"             "JJ"                   "SCH_SSCLASSES_IND"
[10] "SCH_SSCLASSES_ALGG_M" "SCH_SSCLASSES_ALGG_F" "TOT_SSCLASSES_ALGG"
[13] "SCH_SSCLASSES_OTHM_M" "SCH_SSCLASSES_OTHM_F" "TOT_SSCLASSES_OTHM"
[16] "SCH_SSCLASSES_SCI_M"  "SCH_SSCLASSES_SCI_F"  "TOT_SSCLASSES_SCI"
[19] "SCH_SSCLASSES_ENGL_M" "SCH_SSCLASSES_ENGL_F" "TOT_SSCLASSES_ENGL"
[22] "SCH_SSCLASSES_OTHA_M" "SCH_SSCLASSES_OTHA_F" "TOT_SSCLASSES_OTHA"
[25] "TOT_SSCLASSES"
```

8. Merge files together

Here, I planned to merge two files together and keep all observations even if there were no perfect matches, so I used the function full_join(data1,data2, by = "surrogate key"), which retains all observations from both datasets based on the common surrogate key. COMBOKEY, the "surrogate key", in both datasets is unique and has a one-to-one relationship between the datasets for accurate merging. Also, I used colnames() to check the newdata.

```
full_join(adata,cdata1, by = "COMBOKEY")->acdata_full
colnames(acdata_full)
```

```
 [1] "LEA_STATE.x"          "LEA_STATE_NAME.x"     "LEAID.x"
 [4] "LEA_NAME.x"           "SCHID.x"              "SCH_NAME.x"
 [7] "COMBOKEY"             "JJ.x"                 "SCH_SSATHLETICS_IND"
[10] "SCH_SSSPORTS_M"       "SCH_SSSPORTS_F"       "TOT_SSSPORTS"
[13] "SCH_SSTEAMS_M"        "SCH_SSTEAMS_F"        "TOT_SSTEAMS"
[16] "SCH_SSPART_M"         "SCH_SSPART_F"         "TOT_SSPART"
```

```
[19] "MIN-MAX"              "LEA_STATE.y"          "LEA_STATE_NAME.y"
[22] "LEAID.y"              "LEA_NAME.y"           "SCHID.y"
[25] "SCH_NAME.y"           "JJ.y"                 "SCH_SSCLASSES_IND"
[28] "SCH_SSCLASSES_ALGG_M" "SCH_SSCLASSES_ALGG_F" "TOT_SSCLASSES_ALGG"
[31] "SCH_SSCLASSES_OTHM_M" "SCH_SSCLASSES_OTHM_F" "TOT_SSCLASSES_OTHM"
[34] "SCH_SSCLASSES_SCI_M"  "SCH_SSCLASSES_SCI_F"  "TOT_SSCLASSES_SCI"
[37] "SCH_SSCLASSES_ENGL_M" "SCH_SSCLASSES_ENGL_F" "TOT_SSCLASSES_ENGL"
[40] "SCH_SSCLASSES_OTHA_M" "SCH_SSCLASSES_OTHA_F" "TOT_SSCLASSES_OTHA"
[43] "TOT_SSCLASSES"
```

9. Scatterplot

Here, I aimed to determine the correlation between the total number of participants in interscholastic activities and the total number of classes. My hypothesis was Schools with more academic classes would have fewer interscholastic athletic participants since students spent less time on these activities due to higher academic burdens. Using the combined dataset created earlier, I selected Total Participants as the dependent variable (Y) and Total Classes as the independent variable (X), and applied appropriate scaling.ggplot(data)+ geom_point(mapping=aes(x= , y=))

Furthermore, I utilized the labs function to label the title and X & Y axes for better clarity. Based on the graph, it seems that there is no significant relationship between the two variables.

Data Visualization Principles Adherence:

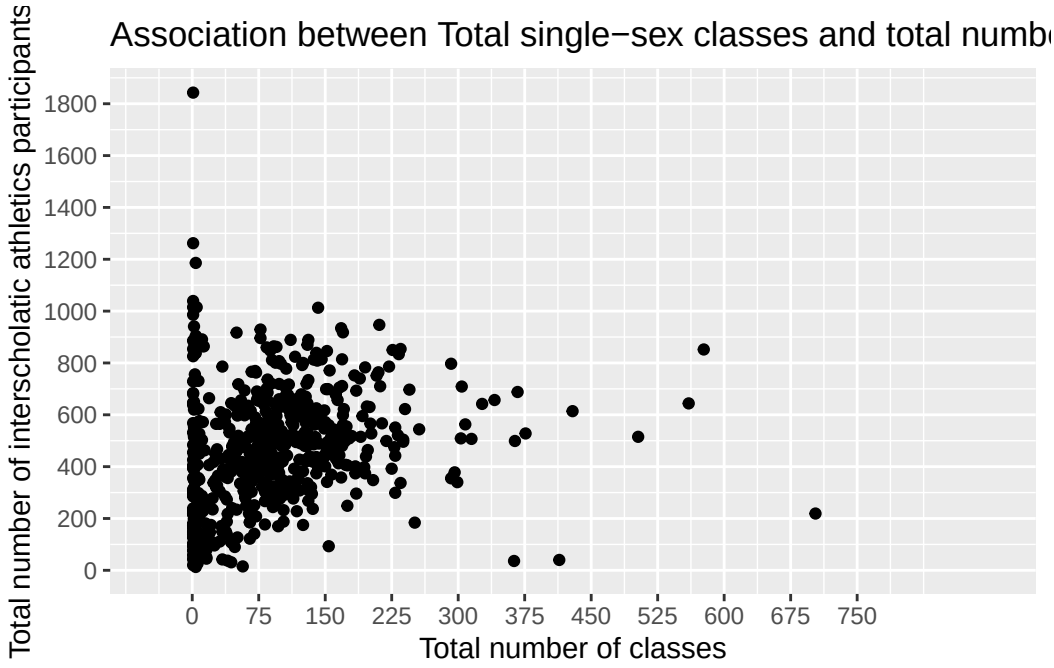
- 1). Proportional Size. Here, the scaling of both the X and Y axes accurately reflects the data.

- 2). Add labels and explanations. Here, I labeled for both the X and Y axes and a title to provide context and overcome potential ambiguity.

- 3). Maximise data-ink ratio. Here, the ink used to represent data rather than extraneous decoration.

```
library(ggplot2)
ggplot(data=acdata_full) +
  geom_point(mapping = aes(x=TOT_SSCLASSES, y=TOT_SSPART))+
  scale_x_continuous(breaks = seq(0,750, by = 75)) +
  scale_y_continuous(breaks = seq(0,2000, by = 200)) +
  labs(
    title = paste("Association between Total single-sex classes and total number of inters
    x = paste("Total number of classes"),
    y = paste("Total number of interscholatic athletics participants"))
```

```
Warning: Removed 2004 rows containing missing values (`geom_point()`).
```



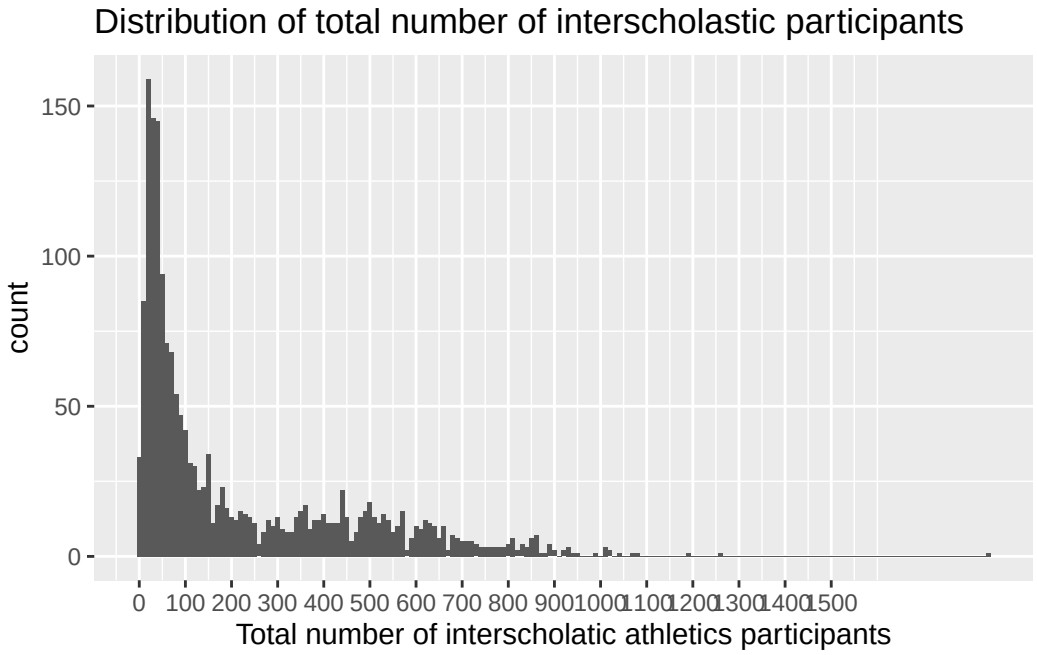
10. Histogram to present distribution

Here, I was interested in understanding the distribution of interscholastic athletic participants since it is considered as an outcome variable. To get a clear picture of the current status of interscholastic athletic participants, I used the geom_histogram function by ggplot packages ggplot(data)+ geom_histogram(mapping=aes(x=)) to present the distribution. I also utilized the lab() to add a title and name the x-axis. The histogram displayed a skewed right distribution shape.

```
ggplot(data=acdata_full) +
  geom_histogram(mapping=aes(x = TOT_SSPART), binwidth=10)+
  scale_x_continuous(breaks = seq(0,1500, by = 100)) +
  labs(
    title = paste("Distribution of total number of interscholastic participants"),
```

```
    x = paste("Total number of interscholatic athletics participants"))
```

```
Warning: Removed 849 rows containing non-finite values (`stat_bin()`).
```



11. Descriptive statistics

Here, I used the describe(data) function to run descriptive statistics. I was surprised to find out that:

1) the minimum number of total classes (TOT_SSCLASSES) was -45 (what is the meaning of the negative value in the original dataset? It seems unusual and confusing.);

2) the total number of interscholastic participants (TOT_SSPART) varied greatly from 1 to 1843; and

3) All mean number of classes (cdata) for males only was higher than that for females.

```
print(describe(acdata_full))
```

	vars	n	mean	sd	median	trimmed	mad
LEA_STATE.x*	1	1790	20.56	13.27	18.0	19.90	13.34
LEA_STATE_NAME.x*	2	1790	20.61	13.28	18.5	19.94	14.08
LEAID.x*	3	1790	509.65	388.64	473.5	486.50	561.91
LEA_NAME.x*	4	1790	602.29	350.74	614.0	599.30	445.52
SCHID.x	5	1790	3545.78	11367.04	1357.5	1812.82	1466.29
SCH_NAME.x*	6	1790	890.40	513.98	890.5	890.42	660.50
COMBOKEY	7	2639	1088202.66	99157.79	1001319.0	1085122.25	1712.40
JJ.x*	8	1790	1.00	0.00	1.0	1.00	0.00
SCH_SSATHLETICS_IND*	9	1790	1.00	0.00	1.0	1.00	0.00
SCH_SSSPORTS_M	10	1790	4.12	4.37	1.0	3.45	1.48
SCH_SSSPORTS_F	11	1790	4.26	4.37	2.0	3.58	1.48
TOT_SSSPORTS	12	1790	8.39	8.64	3.0	7.06	1.48
SCH_SSTEAMS_M	13	1790	6.49	6.84	3.0	5.41	2.97
SCH_SSTEAMS_F	14	1790	6.60	6.72	3.0	5.58	2.97
TOT_SSTEAMS	15	1790	13.09	13.43	6.0	11.02	5.93
SCH_SSPART_M	16	1790	119.55	136.91	51.0	96.04	62.27
SCH_SSPART_F	17	1790	96.26	109.21	42.0	77.57	47.44
TOT_SSPART	18	1790	215.81	242.35	94.0	174.56	108.23
MIN-MAX*	19	1790	321.84	230.74	239.0	295.66	164.57
LEA_STATE.y*	20	1484	14.15	10.72	9.0	12.02	0.00
LEA_STATE_NAME.y*	21	1484	14.15	10.75	9.0	12.03	0.00
LEAID.y*	22	1484	114.59	122.47	59.0	92.71	54.86
LEA_NAME.y*	23	1484	201.81	123.14	198.5	197.89	163.83

SCHID.y	24	1484	4195.16	8536.90	2606.0	3251.35	2615.31
SCH_NAME.y*	25	1484	734.90	425.79	734.5	734.65	547.82
JJ.y*	26	1484	1.04	0.19	1.0	1.00	0.00
SCH_SSCLASSES_IND*	27	1484	1.00	0.00	1.0	1.00	0.00
SCH_SSCLASSES_ALGG_M	28	1484	2.42	5.48	1.0	1.33	1.48
SCH_SSCLASSES_ALGG_F	29	1484	1.38	3.66	0.0	0.65	0.00
TOT_SSCLASSES_ALGG	30	1484	3.80	8.53	1.0	2.15	1.48
SCH_SSCLASSES_OTHM_M	31	1484	1.51	3.21	0.0	0.80	0.00
SCH_SSCLASSES_OTHM_F	32	1484	0.89	2.22	0.0	0.40	0.00
TOT_SSCLASSES_OTHM	33	1484	2.41	5.04	1.0	1.29	1.48
SCH_SSCLASSES_SCI_M	34	1484	2.60	5.01	1.0	1.60	1.48
SCH_SSCLASSES_SCI_F	35	1484	1.63	3.99	0.0	0.86	0.00
TOT_SSCLASSES_SCI	36	1484	4.24	8.39	2.0	2.65	2.97
SCH_SSCLASSES_ENGL_M	37	1484	4.29	8.23	2.0	2.53	2.97
SCH_SSCLASSES_ENGL_F	38	1484	2.46	5.50	1.0	1.28	1.48
TOT_SSCLASSES_ENGL	39	1484	6.76	12.75	3.0	4.06	4.45
SCH_SSCLASSES_OTHA_M	40	1484	23.63	31.61	11.0	17.72	14.83
SCH_SSCLASSES_OTHA_F	41	1484	18.76	26.09	7.0	13.56	10.38
TOT_SSCLASSES_OTHA	42	1484	42.40	55.16	19.0	32.18	26.69
TOT_SSCLASSES	43	1484	59.60	76.78	32.0	45.66	43.00

	min	max	range	skew	kurtosis	se
LEA_STATE.x*	1	48	47	0.35	-1.27	0.31
LEA_STATE_NAME.x*	1	48	47	0.34	-1.27	0.31
LEAID.x*	1	1243	1242	0.31	-1.26	9.19
LEA_NAME.x*	1	1238	1237	0.01	-1.20	8.29
SCHID.x	1	99999	99998	7.87	63.58	268.67

SCH_NAME.x*	1	1780	1779	0.00	-1.20	12.15
COMBOKEY	1000000	1201154	201154	0.25	-1.94	1930.22
JJ.x*	1	1	0	NaN	NaN	0.00
SCH_SSATHLETICS_IND*	1	1	0	NaN	NaN	0.00
SCH_SSSPORTS_M	0	15	15	1.04	-0.50	0.10
SCH_SSSPORTS_F	0	17	17	1.06	-0.41	0.10
TOT_SSSPORTS	1	31	30	1.07	-0.47	0.20
SCH_SSTEAMS_M	0	42	42	1.21	0.83	0.16
SCH_SSTEAMS_F	0	54	54	1.24	1.48	0.16
TOT_SSTEAMS	0	96	96	1.22	1.03	0.32
SCH_SSPART_M	0	863	863	1.37	1.34	3.24
SCH_SSPART_F	0	980	980	1.54	3.05	2.58
TOT_SSPART	1	1843	1842	1.41	1.76	5.73
MIN-MAX*	1	888	887	0.92	-0.30	5.45
LEA_STATE.y*	1	47	46	1.60	1.12	0.28
LEA_STATE_NAME.y*	1	47	46	1.60	1.11	0.28
LEAID.y*	1	451	450	1.32	0.40	3.18
LEA_NAME.y*	1	449	448	0.15	-1.23	3.20
SCHID.y	1	99999	99998	9.59	104.09	221.61
SCH_NAME.y*	1	1472	1471	0.00	-1.21	11.05
JJ.y*	1	2	1	5.00	23.00	0.00
SCH_SSCLASSES_IND*	1	1	0	NaN	NaN	0.00
SCH_SSCLASSES_ALGG_M	-9	85	94	7.10	80.04	0.14
SCH_SSCLASSES_ALGG_F	-9	82	91	10.01	175.12	0.10
TOT_SSCLASSES_ALGG	-9	166	175	8.38	118.23	0.22
SCH_SSCLASSES_OTHM_M	-9	45	54	4.81	37.36	0.08

SCH_SSCLASSES_OTHM_F	-9	25	34	5.12	38.41	0.06
TOT_SSCLASSES_OTHM	-9	66	75	4.90	36.42	0.13
SCH_SSCLASSES_SCI_M	-9	84	93	6.15	67.90	0.13
SCH_SSCLASSES_SCI_F	-9	79	88	8.92	131.29	0.10
TOT_SSCLASSES_SCI	-9	163	172	8.10	115.44	0.22
SCH_SSCLASSES_ENGL_M	-9	100	109	4.63	30.42	0.21
SCH_SSCLASSES_ENGL_F	-9	79	88	5.85	51.56	0.14
TOT_SSCLASSES_ENGL	-9	135	144	4.77	31.49	0.33
SCH_SSCLASSES_OTHA_M	-9	352	361	2.79	14.01	0.82
SCH_SSCLASSES_OTHA_F	-9	211	220	2.37	7.82	0.68
TOT_SSCLASSES_OTHA	-9	523	532	2.51	10.33	1.43
TOT_SSCLASSES	-45	904	949	3.24	19.48	1.99

```
url = "https://github.com/yrosseel/lavaan/raw/master/R/00class.R"
# Download data
download_data <- rio::import(url, format = "csv")
```

```
Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, : Found
and resolved improper quoting in first 100 rows. If the fields are not quoted
(e.g. field separator does not appear within any field), try quote="" to avoid
this warning.
```

```
Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :
Detected 3 column names but the data has 2 columns. Filling rows automatically.
Set fill=TRUE explicitly to avoid this warning.
```

```r
# Show variable names
names(download_data)
```

```
[1] "data.type          = \"character\"" "# \"full\""
[3] "moment\" or \"none"
```

```r
# use describe() function to get descriptive statistics
library(psych)
describe(download_data)
```

```
                                vars   n  mean    sd median trimmed   mad min
data.type          = "character"*    1 242 83.87 61.11   83.5   82.29 83.03   1
# "full"*                         2 242 21.95 30.68    1.0   16.32  0.00   1
moment" or "none*                  3 242  1.02  0.24    1.0    1.00  0.00   1
                                max range  skew kurtosis   se
data.type          = "character"* 191   190  0.09    -1.34 3.93
# "full"*                         99    98  1.17    -0.10 1.97
moment" or "none*                  4     3 10.49   115.19 0.02
```