Tidy Tuesday

Week 37

AUTHOR
Cristian T

PUBLISHED
September 17, 2024

This week we are exploring Economic Diversity and Student Outcomes!

College students are back on campus in the US, so we're exploring economic diversity and student outcomes! The dataset this week comes from Opportunity Insights via an article and associated interactive visualization from the Upshot at the New York Times.

"A new study, based on millions of anonymous tax records, shows that some colleges are even more economically segregated than previously understood, while others are associated with income mobility."

This dataset offers an opportunity to explore the three rules that make a dataset "tidy":

- Each variable is a column; each column is a variable.
- Each observation is a row; each row is an observation.
- Each value is a cell; each cell is a single value.

`college_admissions.csv

variable	class	description
super_opeid	double	Institution OPEID / Cluster ID when combining multiple OPEIDs.
name	character	Name of college (or college group).
par _income_bin	double	Parent household income group based on percentile in the income distribution.
par _income_lab	character	Parent household income label.
attend	double	Test-score-reweighted absolute attendance rate: Calculated as the fraction of students attending that college among all test-takers within a parent income bin in the Pipeline Analysis Sample.
st derr_attend	double	Standard error on the attend variable.
a ttend_level	double	The school average estimates reweighting on test score. Divide the test-score-reweighted absolute variables by this average to calculate the test-score-reweighted relative variables.
attend_sat	double	Absolute attendance rate for specific test score band based on school tier/category.
stderr _attend_sat	double	Standard error on the attend_sat variable.

variable	class	description
atten d_level_sat	double	The school average estimates reweighting on test score. Divide the test-score-reweighted absolute variables by this average to calculate the test-score-reweighted relative variables.
rel_apply	double	Test-score-reweighted relative application rate: Calculated using adjusted score-sending rates, the relative fraction of all standardized test takers who send test scores to a given college.
stder r_rel_apply	double	Standard error on the rel_apply variable.
rel_attend	double	Test-score-reweighted relative attendance rate: Calculated as the fraction of students attending that college among all test-takers within a parent income bin in the Pipeline Analysis Sample. Relative attendance rates are reported as a proportion of the mean attendance rate across all parent income bins for each college.
stderr _rel_attend	double	Standard error on the rel_attend variable.
rel_a tt_cond_app	double	Calculated as the ratio of rel_attend to rel_apply.
re l_apply_sat	double	Relative application rate for specific test score band based on school tier/category. Selected test score band is the 50-point band that had the most attendees in each school tier/category. The selected range: Ivy Plus: SAT 1460-1510; Elite Public: SAT 1180-1230; Top Private: SAT 1410-1460; NESCAC: SAT 1370-1420; Tier 2 Private: SAT 1290-1340; Top 100 Private: SAT 1170-1220; Top 100 Public: SAT 1110-1160; Other Flagship: SAT 1070-1120
stderr_re l_apply_sat	double	Standard error on the rel_apply_sat variable.
rel _attend_sat	double	Relative attendance rate for specific test score band based on school tier/category.
stderr_rel _attend_sat	double	Standard error on the rel_attend_sat variable.
rel_att_c ond_app_sat	double	Relative attendance rate, conditional on application, for specific test score band based on school tier/category
att end_instate	double	Test-score-reweighted absolute attendance rate for in-state students. Only available for public schools.
stderr_att end_instate	double	Standard error on the attend_instate variable.

variable	class	description
attend_level_instate	double	The school average estimates reweighting on test score. Divide the test-score-reweighted absolute variables by this average to calculate the test-score-reweighted relative variables.
attend_instate_sat	double	Absolute estimates on a specific test score for in-state students. Only available for public schools.
std err_attend_instate_sat	double	Standard error on the attend_instate_sat variable.
at tend_level_instate_sat	double	Absolute estimates on a specific test score for in-state students. Only available for public schools.
att end_oostate	double	Test-score-reweighted absolute attendance rate for out-of-state students. Only available for public schools.
stderr_attend_oostate	double	Standard error on the attend_oostate variable.
attend_level_oostate	double	The school average estimates reweighting on test score. Divide the test-score-reweighted absolute variables by this average to calculate the test-score-reweighted relative variables.
attend_oostate_sat	double	Absolute estimates on a specific test score for out-of-state students. Only available for public schools.
std err_attend_oostate_sat	double	Standard error on the attend_oostate_sat variable.
at tend_level_oostate_sat	double	Absolute estimates on a specific test score for out-of-state students. Only available for public schools.
rel_ap ply_instate	double	Test-score-reweighted relative application rate for in-state students. In-state status is measured using the students' address when they take a standardized test. Only available for public schools.
st derr_rel_apply_instate	double	Standard error on the rel_apply_instate variable.
rel_att end_instate	double	Test-score-reweighted relative attendance rate for in-state students. Only available for public schools.
std err_rel_attend_instate	double	Standard error on the rel_attend_instate variable.
re l_att_cond_app_instate	double	Test-score-reweighted relative attendance rate, conditional on application, for in-state students. Only available for public schools.

variable	class	description
rel_ap ply_oostate	double	Test-score-reweighted relative application rate for out-of-state students. In-state status is measured using the students' address when they take a standardized test. Only available for public schools.
st derr_rel_ap ply_oostate	double	Standard error on the rel_apply_oostate variable.
rel_att end_oostate	double	Test-score-reweighted relative attendance rate for out-of-state students. Only available for public schools.
std err_rel_att end_oostate	double	Standard error on the rel_attend_oostate variable.
re l_att_cond_ app_oostate	double	Test-score-reweighted relative attendance rate, conditional on application, for out-of-state students. Only available for public schools.
rel_apply_ instate_sat	double	Relative estimates on a specific test score for in-state students. Only available for public schools.
stderr *rel_apply* instate_sat	double	Standard error on the rel_apply_instate_sat variable.
rel_attend_ instate_sat	double	Relative estimates on a specific test score for in-state students. Only available for public schools.
stderr_ rel_attend_ instate_sat	double	Standard error on the rel_attend_instate_sat variable.
rel_at t_cond_app_ instate_sat	double	Estimates on a specific test score for in-state students. Only available for public schools.
rel_apply_ oostate_sat	double	Relative estimates on a specific test score for out-of-state students. Only available for public schools.
stderr *rel_apply* oostate_sat	double	Standard error on the rel_apply_oostate_sat variable.
rel_attend_ oostate_sat	double	Relative estimates on a specific test score for out-of-state students. Only available for public schools.
stderr_ rel_attend_ oostate_sat	double	Standard error on the rel_attend_oostate_sat variable.

variable	class	description
rel_at t_cond_app_ oostate_sat	double	Estimates on a specific test score for out-of-state students. Only available for public schools.
a ttend_unwgt	double	Unweighted absolute attendance rate: Calculated as the fraction of students attending that college among all test-takers within a parent income bin in the Pipeline Analysis Sample.
stderr_a ttend_unwgt	double	Standard error on the attend_unwgt variable.
attend_ unwgt_level	double	The unweighted school average estimates. Divide the unweighted absolute variables by this average to calculate the unweighted relative variables.
attend_un wgt_instate	double	Unweighted absolute estimates for instate students. Only available for public schools.
stder r_attend_un wgt_instate	double	Standard error on the attend_unwgt_instate variable.
attend_un wgt_oostate	double	Unweighted absolute estimates for out-of-state students. Only available for public schools.
stder r_attend_un wgt_oostate	double	Standard error on the attend_unwgt_oostate variable.
atte nd_unwgt_le vel_instate	double	The unweighted school average estimates. Divide the unweighted absolute variables by this average to calculate the unweighted relative variables.
atte nd_unwgt_le vel_oostate	double	The unweighted school average estimates. Divide the unweighted absolute variables by this average to calculate the unweighted relative variables.
rel_a ttend_unwgt	double	Unweighted relative attendance rate: Calculated as the fraction of students attending that college among all test-takers within a parent income bin in the Pipeline Analysis Sample. Relative attendance rates are reported as a proportion of the mean attendance rate across all parent income bins for each college.
rel_ apply_unwgt	double	Unweighted relative application rate: Calculated using adjusted score-sending rates, the relative fraction of all standardized test takers who send test scores to a given college.
s tderr_rel_a ttend_unwgt	double	Standard error on the rel_attend_unwgt variable.

variable	class	description
stderr_rel_apply_unwgt	double	Standard error on the rel_apply_unwgt variable.
rel_att_cond_app_unwgt	double	Calculated as the ratio of rel_attend_unwgt to rel_apply_unwgt.
rel_attend_unwgt_instate	double	Unweighted relative estimates for instate students. Only available for public schools.
rel_attend_unwgt_oostate	double	Unweighted relative estimates for out-of-state students. Only available for public schools.
stderr_rel_attend_unwgt_instate	double	Standard error on the rel_attend_unwgt_instate variable.
stderr_rel_attend_unwgt_oostate	double	Standard error on the rel_attend_unwgt_oostate variable.
rel_apply_unwgt_instate	double	Unweighted relative estimates for instate students. Only available for public schools.
rel_apply_unwgt_oostate	double	Unweighted relative estimates for out-of-state students. Only available for public schools.
stderr_rel_apply_unwgt_instate	double	Standard error on the rel_apply_unwgt_instate variable.
stderr_rel_apply_unwgt_oostate	double	Standard error on the rel_apply_unwgt_oostate variable.
rel_att_cond_app_unwgt_instate	double	Unweighted estimates for instate students. Only available for public schools.
rel_att_cond_app_unwgt_oostate	double	Unweighted estimates for out-of-state students. Only available for public schools.
public	logical	Indicator for public universities.
flagship	logical	Indicator for public flagship universities (defined using the College Board Annual Survey of Colleges, 2016).
tier	character	Selectivity and type combination: Ivy-Plus (Ivy League colleges plus Stanford, Chicago, Duke, and MIT); Other elite college

variable	class	description
		(Barron's top selectivity category, other than the Ivy-plus, both public and private combined); Highly selective public college (Barron's 2nd selectivity group); Highly selective private college (Barron's 2nd selectivity group); Selective public college (Barron's 3rd, 4th, and 5th selectivity groups); Selective private college (Barron's 3rd, 4th, and 5th selectivity groups) See Chetty, Friedman, Saez, Turner, and Yagan (2020) for more information on how the tier is defined.
tes t_band_tier	character	School group for the test-score band statistics.

Load the data

```
# Load the tidytuesday package
suppressMessages(library(tidytuesdayR)) # For accessing TidyTuesday datasets
suppressMessages(library(skimr)) # For summary and descriptive statistics
suppressMessages(library(tidyverse)) # For data manipulation and visualization
suppressMessages(library(dplyr)) # For data manipulation and transformation
suppressMessages(library(ggplot2)) # For data visualization
suppressMessages(library(RColorBrewer)) # For color palettes in visualizations
suppressMessages(library(ggimage) ) # For adding images to plots


# Load the current week's dataset
tuesdata <- tidytuesdayR::tt_load('2024-09-10')
```

```
Downloading file 1 of 1: `college_admissions.csv`
```

```
# Extract datasets from the TidyTuesday dataset
college_admissions <- tuesdata$college_admissions

# Rename datasets
ca <- college_admissions


# Explore the structure of the dataset
str(ca) # Display the structure of 'college_admissions'
```

```
spc_tbl_ [1,946 × 80] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ super_opeid                   : num [1:1946] 1434 1434 1434 1434 1434 ...
 $ name                          : chr [1:1946] "American University" "American University"
"American University" "American University" ...
 $ par_income_bin                : num [1:1946] 10 30 50 65 75 85 92.5 95.5 96.5 97.5 ...
 $ par_income_lab                : chr [1:1946] "0-20" "20-40" "40-60" "60-70" ...
 $ attend                        : num [1:1946] 0.00112 0.001 0.00141 0.00149 0.0015 ...
 $ stderr_attend                 : num [1:1946] 1.20e-04 9.41e-05 8.24e-05 9.20e-05 7.66e-05
...
 $ attend_level                  : num [1:1946] 0.00161 0.00161 0.00161 0.00161 0.00161 ...
```

```
 $ attend_sat                    : num [1:1946] 0.00136 0.00206 0.00143 0.00141 0.00126 ...
 $ stderr_attend_sat             : num [1:1946] 0.000236 0.000167 0.00015 0.000165 0.000131
...
 $ attend_level_sat              : num [1:1946] 0.00154 0.00154 0.00154 0.00154 0.00154 ...
 $ rel_apply                     : num [1:1946] 0.666 0.681 0.702 0.715 0.722 ...
 $ stderr_rel_apply              : num [1:1946] 0.0265 0.0209 0.0156 0.0174 0.0145 ...
 $ rel_attend                    : num [1:1946] 0.698 0.624 0.875 0.925 0.933 ...
 $ stderr_rel_attend             : num [1:1946] 0.0748 0.0585 0.0512 0.0572 0.0476 ...
 $ rel_att_cond_app              : num [1:1946] 1.047 0.917 1.246 1.293 1.292 ...
 $ rel_apply_sat                 : num [1:1946] 0.732 0.762 0.691 0.655 0.613 ...
 $ stderr_rel_apply_sat          : num [1:1946] 0.0527 0.0421 0.0291 0.0312 0.0254 ...
 $ rel_attend_sat                : num [1:1946] 0.886 1.342 0.931 0.914 0.819 ...
 $ stderr_rel_attend_sat         : num [1:1946] 0.1538 0.1086 0.0978 0.1076 0.0849 ...
 $ rel_att_cond_app_sat          : num [1:1946] 1.21 1.76 1.35 1.4 1.34 ...
 $ attend_instate                : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_attend_instate         : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_level_instate          : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_instate_sat            : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_attend_instate_sat     : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_level_instate_sat      : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_oostate                : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_attend_oostate         : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_level_oostate          : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_oostate_sat            : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_attend_oostate_sat     : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_level_oostate_sat      : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_apply_instate             : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_apply_instate      : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_attend_instate            : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_attend_instate     : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_att_cond_app_instate      : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_apply_oostate             : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_apply_oostate      : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_attend_oostate            : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_attend_oostate     : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_att_cond_app_oostate      : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_apply_instate_sat         : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_apply_instate_sat  : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_attend_instate_sat        : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_attend_instate_sat : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_att_cond_app_instate_sat  : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_apply_oostate_sat         : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_apply_oostate_sat  : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_attend_oostate_sat        : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_attend_oostate_sat : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_att_cond_app_oostate_sat  : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_unwgt                  : num [1:1946] 0.000265 0.000262 0.000407 0.000539 0.000631
...
 $ stderr_attend_unwgt           : num [1:1946] 2.02e-05 1.70e-05 1.86e-05 2.71e-05 2.72e-05
...
 $ attend_unwgt_level            : num [1:1946] 0.000677 0.000677 0.000677 0.000677 0.000677
...
 $ attend_unwgt_instate          : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_attend_unwgt_instate   : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
```

```
 $ attend_unwgt_oostate           : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_attend_unwgt_oostate    : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_unwgt_level_instate     : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ attend_unwgt_level_oostate     : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_attend_unwgt               : num [1:1946] 0.392 0.387 0.601 0.796 0.932 ...
 $ rel_apply_unwgt                : num [1:1946] 0.337 0.393 0.508 0.621 0.723 ...
 $ stderr_rel_attend_unwgt        : num [1:1946] 0.0299 0.0251 0.0274 0.04 0.0401 ...
 $ stderr_rel_apply_unwgt         : num [1:1946] 0.00951 0.00867 0.00864 0.01211 0.01216 ...
 $ rel_att_cond_app_unwgt         : num [1:1946] 1.161 0.985 1.184 1.281 1.289 ...
 $ rel_attend_unwgt_instate       : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_attend_unwgt_oostate       : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_attend_unwgt_instate: num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_attend_unwgt_oostate: num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_apply_unwgt_instate        : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_apply_unwgt_oostate        : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_apply_unwgt_instate : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ stderr_rel_apply_unwgt_oostate : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_att_cond_app_unwgt_instate : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ rel_att_cond_app_unwgt_oostate : num [1:1946] NA NA NA NA NA NA NA NA NA NA ...
 $ public                         : logi [1:1946] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ flagship                       : logi [1:1946] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ tier                           : chr [1:1946] "Highly selective private" "Highly selective
private" "Highly selective private" "Highly selective private" ...
 $ test_band_tier                 : chr [1:1946] "Other Top 100 Private" "Other Top 100
Private" "Other Top 100 Private" "Other Top 100 Private" ...
 - attr(*, "spec")=
  .. cols(
  ..   super_opeid = col_double(),
  ..   name = col_character(),
  ..   par_income_bin = col_double(),
  ..   par_income_lab = col_character(),
  ..   attend = col_double(),
  ..   stderr_attend = col_double(),
  ..   attend_level = col_double(),
  ..   attend_sat = col_double(),
  ..   stderr_attend_sat = col_double(),
  ..   attend_level_sat = col_double(),
  ..   rel_apply = col_double(),
  ..   stderr_rel_apply = col_double(),
  ..   rel_attend = col_double(),
  ..   stderr_rel_attend = col_double(),
  ..   rel_att_cond_app = col_double(),
  ..   rel_apply_sat = col_double(),
  ..   stderr_rel_apply_sat = col_double(),
  ..   rel_attend_sat = col_double(),
  ..   stderr_rel_attend_sat = col_double(),
  ..   rel_att_cond_app_sat = col_double(),
  ..   attend_instate = col_double(),
  ..   stderr_attend_instate = col_double(),
  ..   attend_level_instate = col_double(),
  ..   attend_instate_sat = col_double(),
  ..   stderr_attend_instate_sat = col_double(),
  ..   attend_level_instate_sat = col_double(),
  ..   attend_oostate = col_double(),
```

```
..   stderr_attend_oostate = col_double(),
..   attend_level_oostate = col_double(),
..   attend_oostate_sat = col_double(),
..   stderr_attend_oostate_sat = col_double(),
..   attend_level_oostate_sat = col_double(),
..   rel_apply_instate = col_double(),
..   stderr_rel_apply_instate = col_double(),
..   rel_attend_instate = col_double(),
..   stderr_rel_attend_instate = col_double(),
..   rel_att_cond_app_instate = col_double(),
..   rel_apply_oostate = col_double(),
..   stderr_rel_apply_oostate = col_double(),
..   rel_attend_oostate = col_double(),
..   stderr_rel_attend_oostate = col_double(),
..   rel_att_cond_app_oostate = col_double(),
..   rel_apply_instate_sat = col_double(),
..   stderr_rel_apply_instate_sat = col_double(),
..   rel_attend_instate_sat = col_double(),
..   stderr_rel_attend_instate_sat = col_double(),
..   rel_att_cond_app_instate_sat = col_double(),
..   rel_apply_oostate_sat = col_double(),
..   stderr_rel_apply_oostate_sat = col_double(),
..   rel_attend_oostate_sat = col_double(),
..   stderr_rel_attend_oostate_sat = col_double(),
..   rel_att_cond_app_oostate_sat = col_double(),
..   attend_unwgt = col_double(),
..   stderr_attend_unwgt = col_double(),
..   attend_unwgt_level = col_double(),
..   attend_unwgt_instate = col_double(),
..   stderr_attend_unwgt_instate = col_double(),
..   attend_unwgt_oostate = col_double(),
..   stderr_attend_unwgt_oostate = col_double(),
..   attend_unwgt_level_instate = col_double(),
..   attend_unwgt_level_oostate = col_double(),
..   rel_attend_unwgt = col_double(),
..   rel_apply_unwgt = col_double(),
..   stderr_rel_attend_unwgt = col_double(),
..   stderr_rel_apply_unwgt = col_double(),
..   rel_att_cond_app_unwgt = col_double(),
..   rel_attend_unwgt_instate = col_double(),
..   rel_attend_unwgt_oostate = col_double(),
..   stderr_rel_attend_unwgt_instate = col_double(),
..   stderr_rel_attend_unwgt_oostate = col_double(),
..   rel_apply_unwgt_instate = col_double(),
..   rel_apply_unwgt_oostate = col_double(),
..   stderr_rel_apply_unwgt_instate = col_double(),
..   stderr_rel_apply_unwgt_oostate = col_double(),
..   rel_att_cond_app_unwgt_instate = col_double(),
..   rel_att_cond_app_unwgt_oostate = col_double(),
..   public = col_logical(),
..   flagship = col_logical(),
..   tier = col_character(),
..   test_band_tier = col_character()
```

```
  .. )
 - attr(*, "problems")=<externalptr>
```

```
skim(ca) # Provide detailed summary statistics for 'college_admissions' (missing valu
```

Name	ca
Number of rows	1946
Number of columns	80
_____	
Column type frequency:	
character	4
logical	2
numeric	74
_____	
Group variables	None

Data summary

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
name	0	1	12	49	0	139	0
par_income_lab	0	1	4	7	0	14	0
tier	0	1	8	40	0	6	0
test_band_tier	0	1	6	21	0	6	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
public	0	1	0.37	FAL: 1232, TRU: 714
flagship	0	1	0.21	FAL: 1540, TRU: 406

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	
super_opeid	0	1.00	2528.51	1344.92	108.00	1536.00	2536.00	3223.00	11
par_income_bin	0	1.00	78.17	28.04	10.00	65.00	94.00	98.50	
attend	2	1.00	0.00	0.01	0.00	0.00	0.00	0.01	
stderr_attend	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	
attend_level	0	1.00	0.00	0.00	0.00	0.00	0.00	0.01	
attend_sat	294	0.85	0.00	0.00	0.00	0.00	0.00	0.01	
stderr_attend_sat	278	0.86	0.00	0.00	0.00	0.00	0.00	0.00	
attend_level_sat	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	
rel_apply	0	1.00	1.17	0.56	0.07	0.83	1.05	1.34	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
stderr_rel_apply	0	1.00	0.04	0.04	0.01	0.02	0.03	0.05
rel_attend	2	1.00	1.27	0.93	0.01	0.76	1.03	1.45
stderr_rel_attend	0	1.00	0.11	0.12	0.00	0.05	0.08	0.13
rel_att_cond_app	2	1.00	1.03	0.27	0.05	0.88	1.00	1.15
rel_apply_sat	278	0.86	1.15	0.51	0.19	0.79	1.06	1.36
stderr_rel_apply_sat	278	0.86	0.07	0.05	0.02	0.03	0.06	0.09
rel_attend_sat	294	0.85	1.19	0.75	0.02	0.71	0.99	1.48
stderr_rel_attend_sat	278	0.86	0.18	0.13	0.00	0.08	0.15	0.24
rel_att_cond_app_sat	294	0.85	1.01	0.35	0.04	0.82	0.98	1.17
attend_instate	1334	0.31	0.13	0.08	0.00	0.06	0.13	0.18
stderr_attend_instate	1334	0.31	0.01	0.01	0.00	0.00	0.01	0.01
attend_level_instate	1232	0.37	0.12	0.07	0.02	0.05	0.13	0.17
attend_instate_sat	1334	0.31	0.13	0.10	0.00	0.04	0.11	0.18
stderr_attend_instate_sat	1334	0.31	0.02	0.02	0.00	0.00	0.01	0.02
attend_level_instate_sat	1232	0.37	0.12	0.08	0.01	0.03	0.11	0.16
attend_oostate	1336	0.31	0.00	0.00	0.00	0.00	0.00	0.00
stderr_attend_oostate	1334	0.31	0.00	0.00	0.00	0.00	0.00	0.00
attend_level_oostate	1232	0.37	0.00	0.00	0.00	0.00	0.00	0.00
attend_oostate_sat	1346	0.31	0.00	0.00	0.00	0.00	0.00	0.00
stderr_attend_oostate_sat	1334	0.31	0.00	0.00	0.00	0.00	0.00	0.00
attend_level_oostate_sat	1232	0.37	0.00	0.00	0.00	0.00	0.00	0.00
rel_apply_instate	1334	0.31	1.01	0.17	0.27	0.91	1.00	1.12
stderr_rel_apply_instate	1334	0.31	0.03	0.02	0.01	0.01	0.02	0.03
rel_attend_instate	1334	0.31	1.03	0.30	0.14	0.86	1.01	1.20
stderr_rel_attend_instate	1334	0.31	0.06	0.04	0.01	0.03	0.05	0.07
rel_att_cond_app_instate	1334	0.31	1.01	0.19	0.37	0.90	1.00	1.11
rel_apply_oostate	1334	0.31	1.19	0.55	0.32	0.75	1.10	1.54
stderr_rel_apply_oostate	1334	0.31	0.05	0.04	0.01	0.02	0.03	0.06
rel_attend_oostate	1336	0.31	1.30	0.87	0.24	0.63	1.08	1.74
stderr_rel_attend_oostate	1334	0.31	0.14	0.11	0.00	0.06	0.11	0.18
rel_att_cond_app_oostate	1336	0.31	1.02	0.28	0.34	0.82	1.01	1.18
rel_apply_instate_sat	1334	0.31	1.02	0.20	0.31	0.90	1.00	1.16
stderr_rel_apply_instate_sat	1334	0.31	0.06	0.04	0.02	0.03	0.05	0.07
rel_attend_instate_sat	1334	0.31	1.07	0.44	0.09	0.79	1.00	1.27
stderr_rel_attend_instate_sat	1334	0.31	0.14	0.10	0.03	0.07	0.11	0.16
rel_att_cond_app_instate_sat	1334	0.31	1.03	0.30	0.15	0.86	1.00	1.15
rel_apply_oostate_sat	1334	0.31	1.30	0.70	0.28	0.73	1.12	1.73

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
stderr_rel_apply_oostate_sat	1334	0.31	0.11	0.09	0.02	0.04	0.08	0.15
rel_attend_oostate_sat	1346	0.31	1.49	1.21	0.06	0.55	1.07	2.09
stderr_rel_attend_oostate_sat	1334	0.31	0.34	0.31	0.00	0.11	0.24	0.47
rel_att_cond_app_oostate_sat	1346	0.31	1.05	0.42	0.10	0.76	0.99	1.25
attend_unwgt	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00
stderr_attend_unwgt	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00
attend_unwgt_level	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00
attend_unwgt_instate	1334	0.31	0.09	0.08	0.00	0.03	0.07	0.14
stderr_attend_unwgt_instate	1334	0.31	0.01	0.01	0.00	0.00	0.00	0.01
attend_unwgt_oostate	1337	0.31	0.00	0.00	0.00	0.00	0.00	0.00
stderr_attend_unwgt_oostate	1334	0.31	0.00	0.00	0.00	0.00	0.00	0.00
attend_unwgt_level_instate	1232	0.37	0.07	0.05	0.01	0.02	0.07	0.10
attend_unwgt_level_oostate	1232	0.37	0.00	0.00	0.00	0.00	0.00	0.00
rel_attend_unwgt	1	1.00	2.48	3.10	0.00	0.62	1.41	2.94
rel_apply_unwgt	0	1.00	1.98	1.77	0.09	0.71	1.42	2.62
stderr_rel_attend_unwgt	0	1.00	0.20	0.35	0.00	0.03	0.07	0.22
stderr_rel_apply_unwgt	0	1.00	0.06	0.09	0.00	0.01	0.03	0.07
rel_att_cond_app_unwgt	1	1.00	1.06	0.39	0.01	0.84	1.00	1.20
rel_attend_unwgt_instate	1334	0.31	1.30	0.73	0.18	0.79	1.10	1.64
rel_attend_unwgt_oostate	1337	0.31	1.82	1.67	0.09	0.53	1.29	2.82
stderr_rel_attend_unwgt_instate	1334	0.31	0.07	0.05	0.01	0.03	0.05	0.09
stderr_rel_attend_unwgt_oostate	1334	0.31	0.16	0.15	0.00	0.04	0.10	0.24
rel_apply_unwgt_instate	1334	0.31	1.19	0.43	0.35	0.86	1.10	1.49
rel_apply_unwgt_oostate	1334	0.31	1.57	1.08	0.17	0.69	1.31	2.26
stderr_rel_apply_unwgt_instate	1334	0.31	0.03	0.02	0.00	0.01	0.02	0.04
stderr_rel_apply_unwgt_oostate	1334	0.31	0.05	0.04	0.00	0.01	0.03	0.07
rel_att_cond_app_unwgt_instate	1334	0.31	1.04	0.27	0.40	0.86	1.01	1.17
rel_att_cond_app_unwgt_oostate	1337	0.31	1.02	0.33	0.19	0.78	1.00	1.21

```
# Export data
#write.csv(college_admissions, "college_admissions.csv", row.names = FALSE)



#tidytuesdayR::use_tidytemplate()
```

```
#### Clean the data

# 42 columns have a significant number of missing values (over 1300).
```

EDA

```
# Summarize college_admissions
summary(college_admissions)
```

```
 super_opeid        name           par_income_bin   par_income_lab
Min.   :  108   Length:1946        Min.   : 10.00   Length:1946
1st Qu.: 1536   Class :character   1st Qu.: 65.00   Class :character
Median : 2536   Mode  :character   Median : 94.00   Mode  :character
Mean   : 2529                      Mean   : 78.17
3rd Qu.: 3223                      3rd Qu.: 98.50
Max.   :11649                      Max.   :100.00

    attend          stderr_attend        attend_level         attend_sat
Min.   :0.0000289   Min.   :0.0000000   Min.   :0.0004208   Min.   :0.00002
1st Qu.:0.0014042   1st Qu.:0.0001146   1st Qu.:0.0013405   1st Qu.:0.00108
Median :0.0030317   Median :0.0002052   Median :0.0025648   Median :0.00236
Mean   :0.0048257   Mean   :0.0003537   Mean   :0.0040984   Mean   :0.00401
3rd Qu.:0.0065213   3rd Qu.:0.0003800   3rd Qu.:0.0058981   3rd Qu.:0.00532
Max.   :0.0460693   Max.   :0.0054024   Max.   :0.0201221   Max.   :0.04233
NA's   :2                                                   NA's   :294
stderr_attend_sat attend_level_sat      rel_apply       stderr_rel_apply
Min.   :0.00000   Min.   :0.0003987   Min.   :0.07044   Min.   :0.007235
1st Qu.:0.00017   1st Qu.:0.0010742   1st Qu.:0.83192   1st Qu.:0.018424
Median :0.00031   Median :0.0021369   Median :1.04823   Median :0.027626
Mean   :0.00048   Mean   :0.0035109   Mean   :1.17263   Mean   :0.039336
3rd Qu.:0.00060   3rd Qu.:0.0042152   3rd Qu.:1.33672   3rd Qu.:0.047375
Max.   :0.00375   Max.   :0.0188583   Max.   :5.87360   Max.   :0.529336
NA's   :278
  rel_attend       stderr_rel_attend rel_att_cond_app  rel_apply_sat
Min.   : 0.01001   Min.   :0.00000   Min.   :0.05321   Min.   :0.1882
1st Qu.: 0.75943   1st Qu.:0.04634   1st Qu.:0.87686   1st Qu.:0.7904
Median : 1.02585   Median :0.07896   Median :1.00280   Median :1.0561
Mean   : 1.27325   Mean   :0.11122   Mean   :1.02935   Mean   :1.1462
3rd Qu.: 1.44678   3rd Qu.:0.13243   3rd Qu.:1.14920   3rd Qu.:1.3623
Max.   :10.26102   Max.   :1.57053   Max.   :3.05958   Max.   :4.7022
NA's   :2                            NA's   :2         NA's   :278
stderr_rel_apply_sat rel_attend_sat    stderr_rel_attend_sat
Min.   :0.01530      Min.   :0.01877   Min.   :0.00000
1st Qu.:0.03348      1st Qu.:0.70931   1st Qu.:0.08469
Median :0.05573      Median :0.98758   Median :0.14685
Mean   :0.06879      Mean   :1.19216   Mean   :0.18422
3rd Qu.:0.09026      3rd Qu.:1.47978   3rd Qu.:0.23587
Max.   :0.56117      Max.   :8.03421   Max.   :0.95318
NA's   :278          NA's   :294       NA's   :278
rel_att_cond_app_sat attend_instate   stderr_attend_instate
Min.   :0.04018      Min.   :0.0043   Min.   :0.0004
1st Qu.:0.82093      1st Qu.:0.0550   1st Qu.:0.0024
Median :0.98323      Median :0.1259   Median :0.0054
Mean   :1.01293      Mean   :0.1304   Mean   :0.0080
3rd Qu.:1.16730      3rd Qu.:0.1831   3rd Qu.:0.0096
Max.   :3.84603      Max.   :0.4147   Max.   :0.0773
NA's   :294          NA's   :1334     NA's   :1334
attend_level_instate attend_instate_sat stderr_attend_instate_sat
```

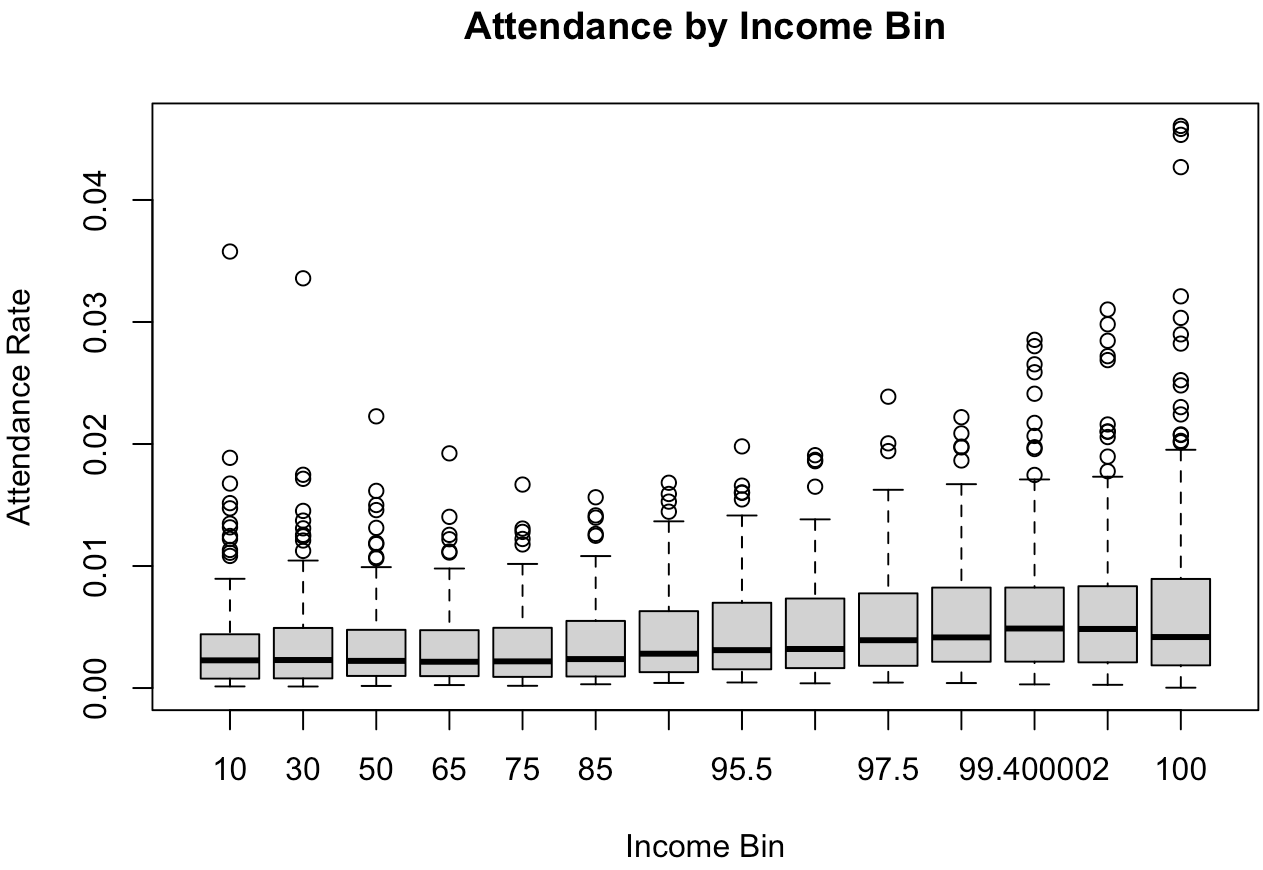
```
Min.   :0.0189       Min.   :0.0016     Min.   :0.0013
1st Qu.:0.0536       1st Qu.:0.0382     1st Qu.:0.0043
Median :0.1322       Median :0.1066     Median :0.0091
Mean   :0.1242       Mean   :0.1287     Mean   :0.0162
3rd Qu.:0.1705       3rd Qu.:0.1838     3rd Qu.:0.0206
Max.   :0.3603       Max.   :0.6567     Max.   :0.1458
NA's   :1232         NA's   :1334       NA's   :1334
attend_level_instate_sat attend_oostate   stderr_attend_oostate
Min.   :0.0143           Min.   :0.0000   Min.   :0e+00
1st Qu.:0.0347           1st Qu.:0.0003   1st Qu.:0e+00
Median :0.1141           Median :0.0007   Median :1e-04
Mean   :0.1181           Mean   :0.0014   Mean   :1e-04
3rd Qu.:0.1579           3rd Qu.:0.0018   3rd Qu.:2e-04
Max.   :0.3881           Max.   :0.0192   Max.   :8e-04
NA's   :1232             NA's   :1336     NA's   :1334
attend_level_oostate attend_oostate_sat stderr_attend_oostate_sat
Min.   :0.0000       Min.   :0.0000     Min.   :0e+00
1st Qu.:0.0004       1st Qu.:0.0002     1st Qu.:1e-04
Median :0.0008       Median :0.0005     Median :1e-04
Mean   :0.0010       Mean   :0.0011     Mean   :2e-04
3rd Qu.:0.0014       3rd Qu.:0.0013     3rd Qu.:2e-04
Max.   :0.0046       Max.   :0.0116     Max.   :9e-04
NA's   :1232         NA's   :1346       NA's   :1334
attend_level_oostate_sat rel_apply_instate stderr_rel_apply_instate
Min.   :0.0000           Min.   :0.2688    Min.   :0.0065
1st Qu.:0.0002           1st Qu.:0.9081    1st Qu.:0.0141
Median :0.0006           Median :1.0003    Median :0.0202
Mean   :0.0007           Mean   :1.0120    Mean   :0.0267
3rd Qu.:0.0011           3rd Qu.:1.1216    3rd Qu.:0.0319
Max.   :0.0018           Max.   :1.6995    Max.   :0.1330
NA's   :1232             NA's   :1334      NA's   :1334
rel_attend_instate stderr_rel_attend_instate rel_att_cond_app_instate
Min.   :0.1358     Min.   :0.0149            Min.   :0.3741
1st Qu.:0.8613     1st Qu.:0.0329            1st Qu.:0.8996
Median :1.0126     Median :0.0463            Median :1.0027
Mean   :1.0314     Mean   :0.0602            Mean   :1.0057
3rd Qu.:1.1971     3rd Qu.:0.0715            3rd Qu.:1.1084
Max.   :2.1925     Max.   :0.3114            Max.   :1.6960
NA's   :1334       NA's   :1334              NA's   :1334
rel_apply_oostate stderr_rel_apply_oostate rel_attend_oostate
Min.   :0.3181    Min.   :0.0085           Min.   :0.2368
1st Qu.:0.7542    1st Qu.:0.0214           1st Qu.:0.6266
Median :1.0958    Median :0.0347           Median :1.0845
Mean   :1.1901    Mean   :0.0456           Mean   :1.2980
3rd Qu.:1.5350    3rd Qu.:0.0585           3rd Qu.:1.7432
Max.   :4.1396    Max.   :0.2622           Max.   :5.8552
NA's   :1334      NA's   :1334             NA's   :1336
stderr_rel_attend_oostate rel_att_cond_app_oostate rel_apply_instate_sat
Min.   :0.0000            Min.   :0.3401           Min.   :0.3081
1st Qu.:0.0584            1st Qu.:0.8223           1st Qu.:0.8957
Median :0.1088            Median :1.0116           Median :1.0002
Mean   :0.1393            Mean   :1.0218           Mean   :1.0229
3rd Qu.:0.1844            3rd Qu.:1.1796           3rd Qu.:1.1578
Max.   :0.7452            Max.   :2.4170           Max.   :1.5694
```

```
NA's   :1334              NA's   :1336            NA's   :1334
stderr_rel_apply_instate_sat rel_attend_instate_sat
Min.   :0.0156               Min.   :0.0909
1st Qu.:0.0292               1st Qu.:0.7904
Median :0.0459               Median :0.9977
Mean   :0.0594               Mean   :1.0727
3rd Qu.:0.0717               3rd Qu.:1.2661
Max.   :0.2791               Max.   :2.9527
NA's   :1334                 NA's   :1334
stderr_rel_attend_instate_sat rel_att_cond_app_instate_sat
Min.   :0.0291                Min.   :0.1480
1st Qu.:0.0674                1st Qu.:0.8619
Median :0.1058                Median :0.9995
Mean   :0.1368                Mean   :1.0288
3rd Qu.:0.1647                3rd Qu.:1.1467
Max.   :0.7127                Max.   :2.1705
NA's   :1334                  NA's   :1334
rel_apply_oostate_sat stderr_rel_apply_oostate_sat rel_attend_oostate_sat
Min.   :0.2802        Min.   :0.0153               Min.   :0.0590
1st Qu.:0.7268        1st Qu.:0.0428               1st Qu.:0.5465
Median :1.1228        Median :0.0832               Median :1.0677
Mean   :1.2956        Mean   :0.1089               Mean   :1.4858
3rd Qu.:1.7318        3rd Qu.:0.1479               3rd Qu.:2.0876
Max.   :4.2104        Max.   :0.6688               Max.   :9.0105
NA's   :1334          NA's   :1334                 NA's   :1346
stderr_rel_attend_oostate_sat rel_att_cond_app_oostate_sat  attend_unwgt
Min.   :0.0000                Min.   :0.0969               Min.   :0.0000018
1st Qu.:0.1123                1st Qu.:0.7625               1st Qu.:0.0003662
Median :0.2429                Median :0.9904               Median :0.0011591
Mean   :0.3385                Mean   :1.0482               Mean   :0.0020019
3rd Qu.:0.4658                3rd Qu.:1.2464               3rd Qu.:0.0026140
Max.   :2.1375                Max.   :2.5915               Max.   :0.0215718
NA's   :1334                  NA's   :1346                 NA's   :1
stderr_attend_unwgt attend_unwgt_level  attend_unwgt_instate
Min.   :0.000e+00   Min.   :9.166e-05   Min.   :0.0038
1st Qu.:2.323e-05   1st Qu.:3.061e-04   1st Qu.:0.0257
Median :6.220e-05   Median :6.328e-04   Median :0.0695
Mean   :1.137e-04   Mean   :1.040e-03   Mean   :0.0934
3rd Qu.:1.457e-04   3rd Qu.:1.581e-03   3rd Qu.:0.1432
Max.   :1.332e-03   Max.   :4.599e-03   Max.   :0.3539
                                        NA's   :1334
stderr_attend_unwgt_instate attend_unwgt_oostate stderr_attend_unwgt_oostate
Min.   :0.0003              Min.   :0.0000       Min.   :0e+00
1st Qu.:0.0011              1st Qu.:0.0001       1st Qu.:0e+00
Median :0.0026              Median :0.0004       Median :0e+00
Mean   :0.0052              Mean   :0.0008       Mean   :1e-04
3rd Qu.:0.0067              3rd Qu.:0.0011       3rd Qu.:1e-04
Max.   :0.0430              Max.   :0.0120       Max.   :3e-04
NA's   :1334                NA's   :1337         NA's   :1334
attend_unwgt_level_instate attend_unwgt_level_oostate rel_attend_unwgt
Min.   :0.0130             Min.   :0.0000             Min.   : 0.003087
1st Qu.:0.0218             1st Qu.:0.0002             1st Qu.: 0.619009
Median :0.0653             Median :0.0003             Median : 1.409683
Mean   :0.0737             Mean   :0.0004             Mean   : 2.483880
```

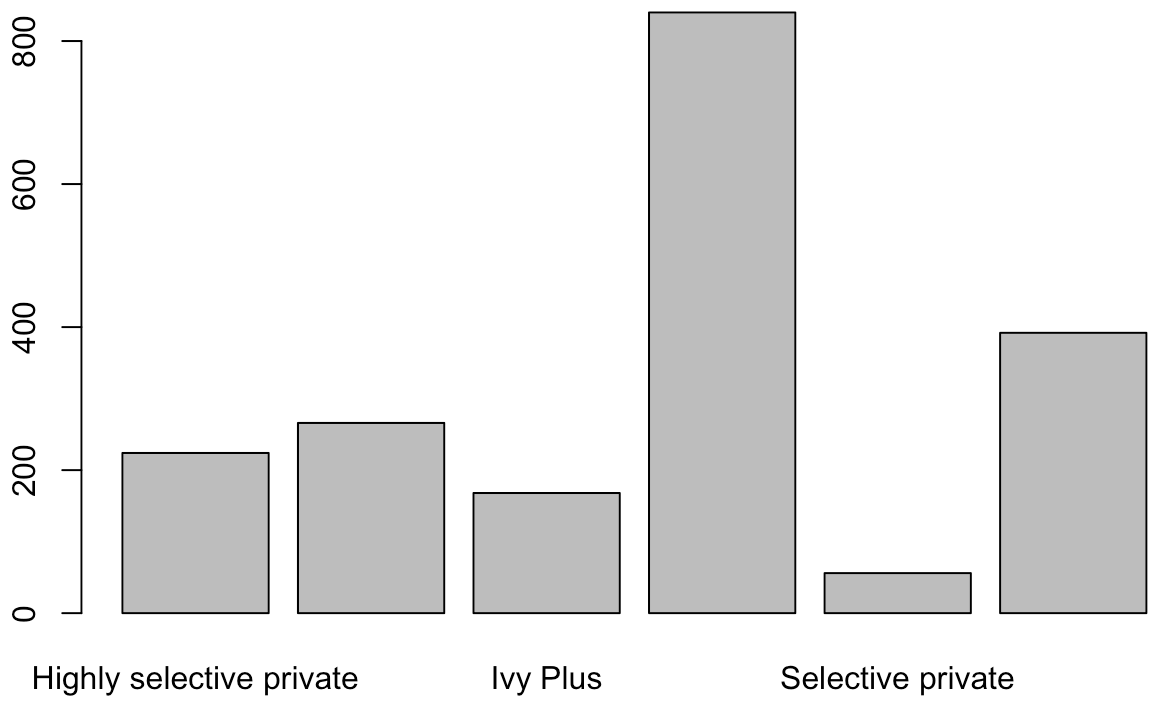
```
3rd Qu.:0.1021            3rd Qu.:0.0007            3rd Qu.: 2.937194
Max.   :0.2387            Max.   :0.0015            Max.   :30.801332
NA's   :1232              NA's   :1232              NA's   :1
rel_apply_unwgt    stderr_rel_attend_unwgt stderr_rel_apply_unwgt
Min.   : 0.09442   Min.   :0.00000         Min.   :0.004941
1st Qu.: 0.70521   1st Qu.:0.03258         1st Qu.:0.011785
Median : 1.42419   Median :0.07486         Median :0.029878
Mean   : 1.98410   Mean   :0.19758         Mean   :0.059705
3rd Qu.: 2.61664   3rd Qu.:0.22436         3rd Qu.:0.066425
Max.   :13.77205   Max.   :3.20504         Max.   :1.056522

rel_att_cond_app_unwgt rel_attend_unwgt_instate rel_attend_unwgt_oostate
Min.   :0.005539       Min.   :0.1776           Min.   : 0.0885
1st Qu.:0.840829       1st Qu.:0.7902           1st Qu.: 0.5315
Median :1.001438       Median :1.1046           Median : 1.2900
Mean   :1.058922       Mean   :1.2962           Mean   : 1.8196
3rd Qu.:1.200485       3rd Qu.:1.6427           3rd Qu.: 2.8219
Max.   :4.948084       Max.   :4.7005           Max.   :12.1050
NA's   :1              NA's   :1334             NA's   :1337
stderr_rel_attend_unwgt_instate stderr_rel_attend_unwgt_oostate
Min.   :0.0088                  Min.   :0.0000
1st Qu.:0.0255                  1st Qu.:0.0412
Median :0.0462                  Median :0.1018
Mean   :0.0660                  Mean   :0.1612
3rd Qu.:0.0883                  3rd Qu.:0.2402
Max.   :0.3365                  Max.   :0.9325
NA's   :1334                    NA's   :1334
rel_apply_unwgt_instate rel_apply_unwgt_oostate stderr_rel_apply_unwgt_instate
Min.   :0.3490          Min.   :0.1730          Min.   :0.0048
1st Qu.:0.8569          1st Qu.:0.6898          1st Qu.:0.0109
Median :1.0987          Median :1.3068          Median :0.0197
Mean   :1.1862          Mean   :1.5723          Mean   :0.0276
3rd Qu.:1.4943          3rd Qu.:2.2621          3rd Qu.:0.0363
Max.   :2.3979          Max.   :7.0723          Max.   :0.1207
NA's   :1334            NA's   :1334            NA's   :1334
stderr_rel_apply_unwgt_oostate rel_att_cond_app_unwgt_instate
Min.   :0.0050                 Min.   :0.3989
1st Qu.:0.0142                 1st Qu.:0.8625
Median :0.0329                 Median :1.0110
Mean   :0.0489                 Mean   :1.0359
3rd Qu.:0.0711                 3rd Qu.:1.1704
Max.   :0.3051                 Max.   :2.0890
NA's   :1334                   NA's   :1334
rel_att_cond_app_unwgt_oostate   public         flagship
Min.   :0.1896                 Mode :logical   Mode :logical
1st Qu.:0.7835                 FALSE:1232      FALSE:1540
Median :0.9986                 TRUE :714       TRUE :406
Mean   :1.0161
3rd Qu.:1.2054
Max.   :2.1428
NA's   :1337
    tier           test_band_tier
Length:1946        Length:1946
Class :character   Class :character
```

```
Mode  :character   Mode  :character
```

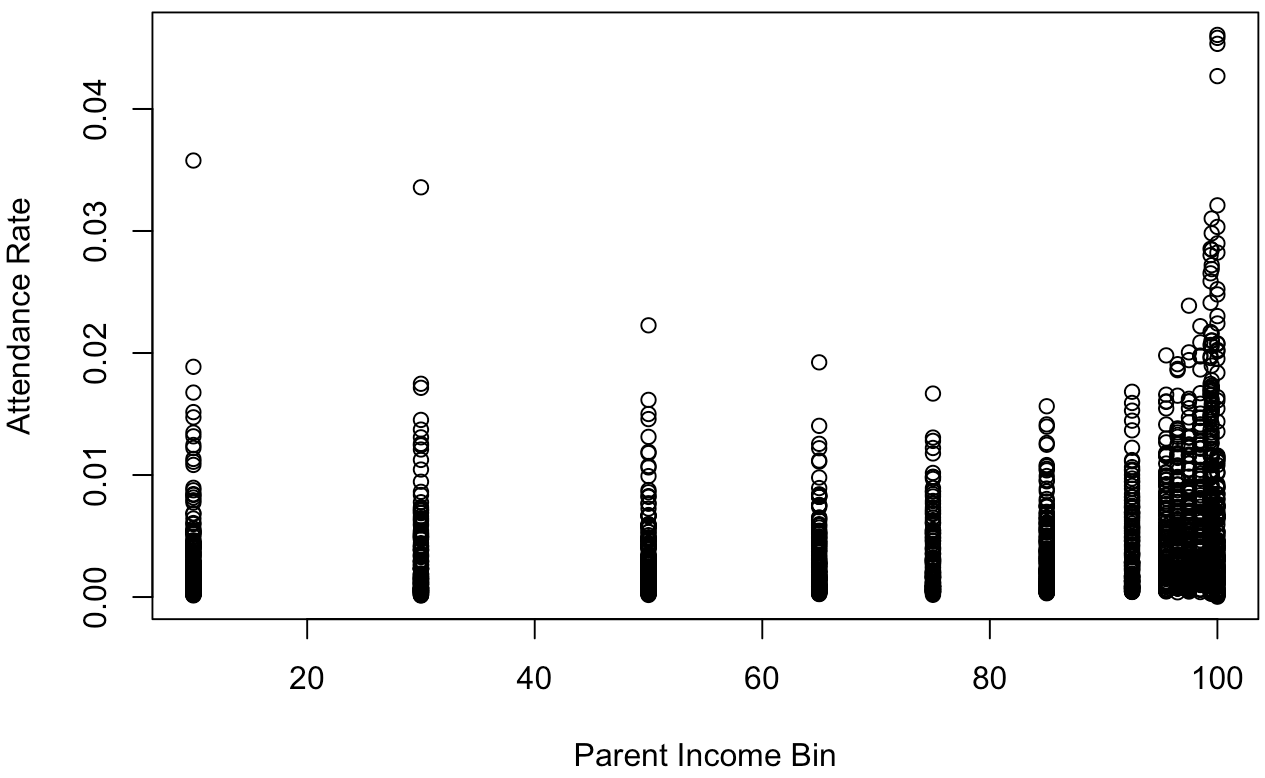
```
# Create a boxplot parent income vs attendance
boxplot(attend ~ par_income_bin, data=ca, main="Attendance by Income Bin", xlab="Inco
```



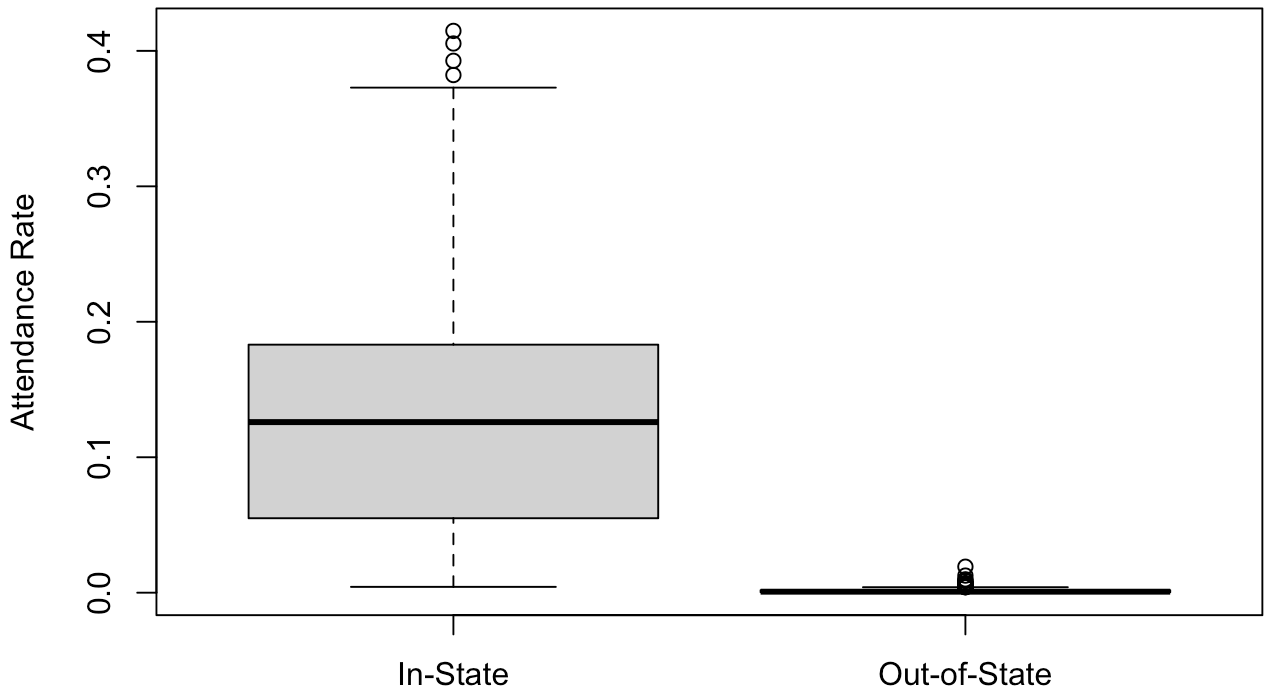

```
# Create a bar plot institutions by their tier
barplot(table(ca$tier), main="Number of Institutions by Tier")
```



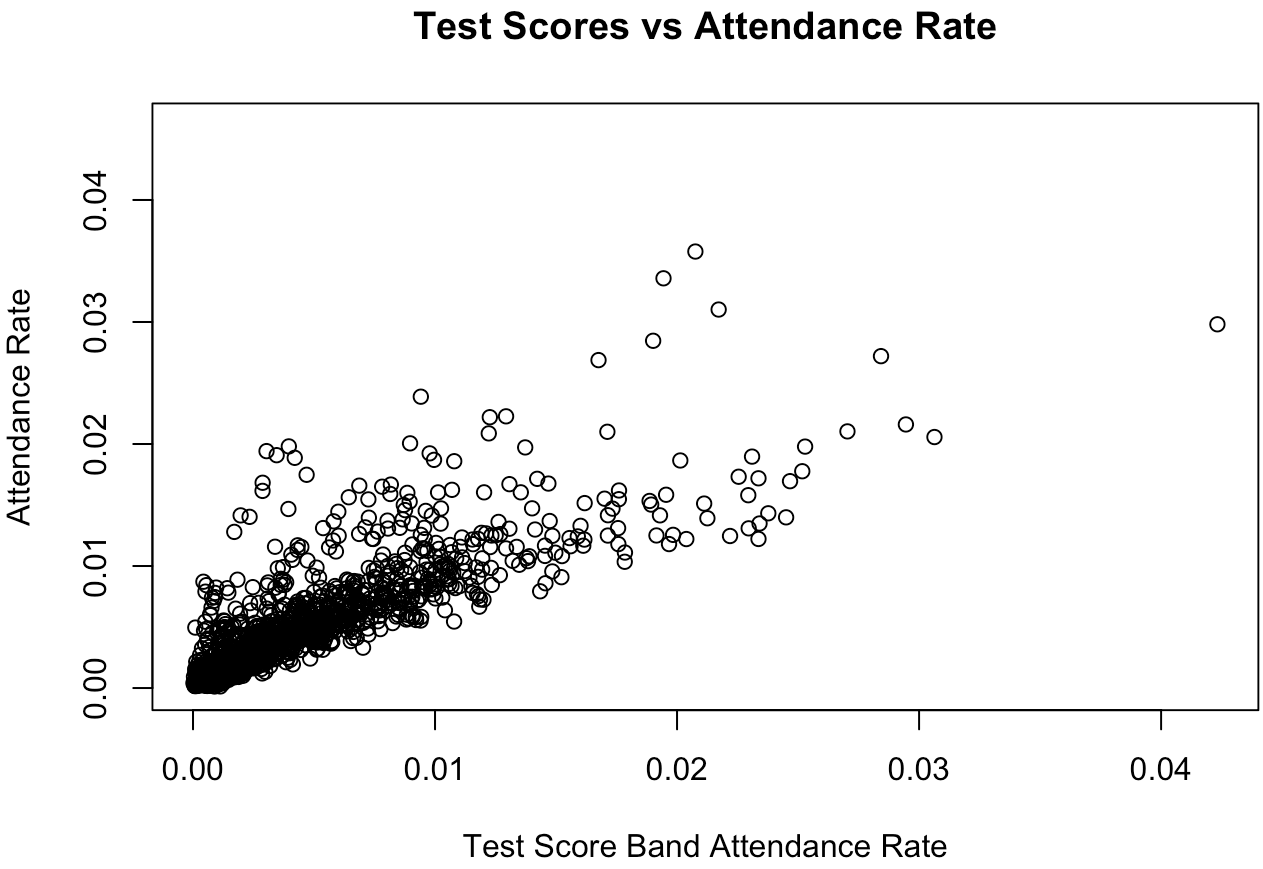

```
# Create a scatterplot parent income v attendance
plot(ca$par_income_bin, ca$attend, main="Parent Income vs Attendance Rate", xlab="Par
```




```
# Create a boxplot in-state vs out-of-state attendance
boxplot(ca$attend_instate, ca$attend_oostate, names=c("In-State", "Out-of-State"), ma
```




```
# Create a scatterplot SAT attendance vs overall attendance rates
plot(ca$attend_sat, ca$attend, main="Test Scores vs Attendance Rate", xlab="Test Scor
```




```r
# Filter higher-income
attend_by_income_bin <- ca %>%
  filter(par_income_bin > 80) %>%
  select(name, par_income_bin, attend, tier) %>% # par_income_lab
  group_by(par_income_bin)
attend_by_income_bin
```

name	par_income_bin	attend
<chr>	<dbl>	<dbl>
American University	85.0	1.524228e-03
American University	92.5	1.695167e-03
American University	95.5	2.261346e-03
American University	96.5	2.109482e-03
American University	97.5	2.267590e-03
American University	98.5	2.484247e-03
American University	99.4	3.078112e-03
American University	99.5	3.030860e-03
American University	100.0	2.873795e-03
Amherst College	85.0	1.348111e-03

1-10 of 1,251 rows | 1-3 of 4 columns

Previous 1 2 3 4 5 6 … 126 Next

```r
# Summarize higher-income by mean, median, and standard deviation
summary_stats <- ca %>%
```

```
  group_by(par_income_bin) %>%
  summarize(
    mean_attend = mean(attend, na.rm = TRUE),
    median_attend = median(attend, na.rm = TRUE),
    sd_attend = sd(attend, na.rm = TRUE)
  )
summary_stats
```

par_income_bin	mean_attend	median_attend	sd_attend
<dbl>	<dbl>	<dbl>	<dbl>
10.0	0.003665678	0.002270875	0.004638523
30.0	0.003656823	0.002295467	0.004411924
50.0	0.003504727	0.002226353	0.003659813
65.0	0.003337591	0.002154475	0.003210391
75.0	0.003412416	0.002186013	0.003180826
85.0	0.003668668	0.002373076	0.003354503
92.5	0.004183249	0.002813860	0.003673880
95.5	0.004668307	0.003105551	0.004072769
96.5	0.004953355	0.003199434	0.004236486
97.5	0.005434125	0.003918635	0.004658010

1-10 of 14 rows Previous **1** 2 Next

```
#  As income increases, both the mean and median attendance rates rise, suggesting th

# Filter data for lower income bins (e.g., income <= 50)
# low_income_data <- ca %>%
#   filter(par_income_bin <= 30)

# Group by school and calculate number of students from lower-income bins attending e
# schools_for_poor_kids <- low_income_data %>%
#   group_by(name) %>%
#   summarise(weighted_attendance_rate = sum(attend, na.rm = TRUE)) %>%
#   arrange(desc(weighted_attendance_rate))
# schools_for_poor_kids

# Filter data for upper income bins (e.g., income >= 92.5)
# rich_income_data <- ca %>%
#   filter(par_income_bin <= 99.5)

# Group by school and calculate number of students from lower-income bins attending e
# schools_for_rich_kids <- rich_income_data %>%
#   group_by(name) %>%
#   summarise(weighted_attendance_rate = sum(attend, na.rm = TRUE)) %>%
#   arrange(desc(weighted_attendance_rate))
# schools_for_rich_kids


# unique(ca$name)

# ND_attend_by_income_bin <- ca %>%
```

```
#   filter(name == "University Of Notre Dame") %>%
#   select(par_income_bin, attend, tier) %>%
#   group_by(par_income_bin)

# attend_by_income_bin <- ca %>%
#   filter(name == "American University") %>%
#   select(par_income_bin, attend, tier) %>%
#   group_by(par_income_bin)

# NM_attend_by_income_bin <- ca %>%
  # filter(name == "University Of New Mexico") %>%
  # select(par_income_bin, attend, tier) %>%
  # group_by(par_income_bin)
# NM_attend_by_income_bin
```

Plot 1 Comparison of Attendance Rates by Income Group

```
# Filter data for lower income bins income <= 30
low_income_data <- ca %>%
  filter(par_income_bin <= 30)

# Group by school and calculate number of students from lower-income bins attending e
schools_for_poor_kids <- low_income_data %>%
  group_by(name) %>% # Group data by school name
  summarise(weighted_attendance_rate = sum(attend, na.rm = TRUE)) %>% # Summarize att
  arrange(desc(weighted_attendance_rate)) # Arrange schools in descending order of at
schools_for_poor_kids
```

name <chr>
Harvard University
University Of California, Berkeley
Yale University
Massachusetts Institute Of Technology
Princeton University
Stanford University
Cornell University
University Of California, Los Angeles
New York University
Columbia University In The City Of New York

1-10 of 139 rows | 1-1 of 2 columns

Previous 1 2 3 4 5 6 ... 14 Next

```
# Filter data for upper income bins income >= 92.5
super_rich_income_data <- ca %>%
  filter(par_income_bin <= 99.5)

# Group by school and calculate number of students from lower-income bins attending e
schools_for_super_rich_kids <- super_rich_income_data %>%
  group_by(name) %>% # Group data by school name
```

```
  summarise(weighted_attendance_rate = sum(attend, na.rm = TRUE)) %>% # Summarize att
  arrange(desc(weighted_attendance_rate)) # Arrange schools in descending order of at
schools_for_super_rich_kids
```

name <chr>
Harvard University
Yale University
University Of California, Berkeley
Princeton University
University Of Michigan - Ann Arbor
Stanford University
Massachusetts Institute Of Technology
University Of Pennsylvania
Cornell University
University Of Chicago

1-10 of 139 rows | 1-1 of 2 columns Previous **1** 2 3 4 5 6 ... 14 Next

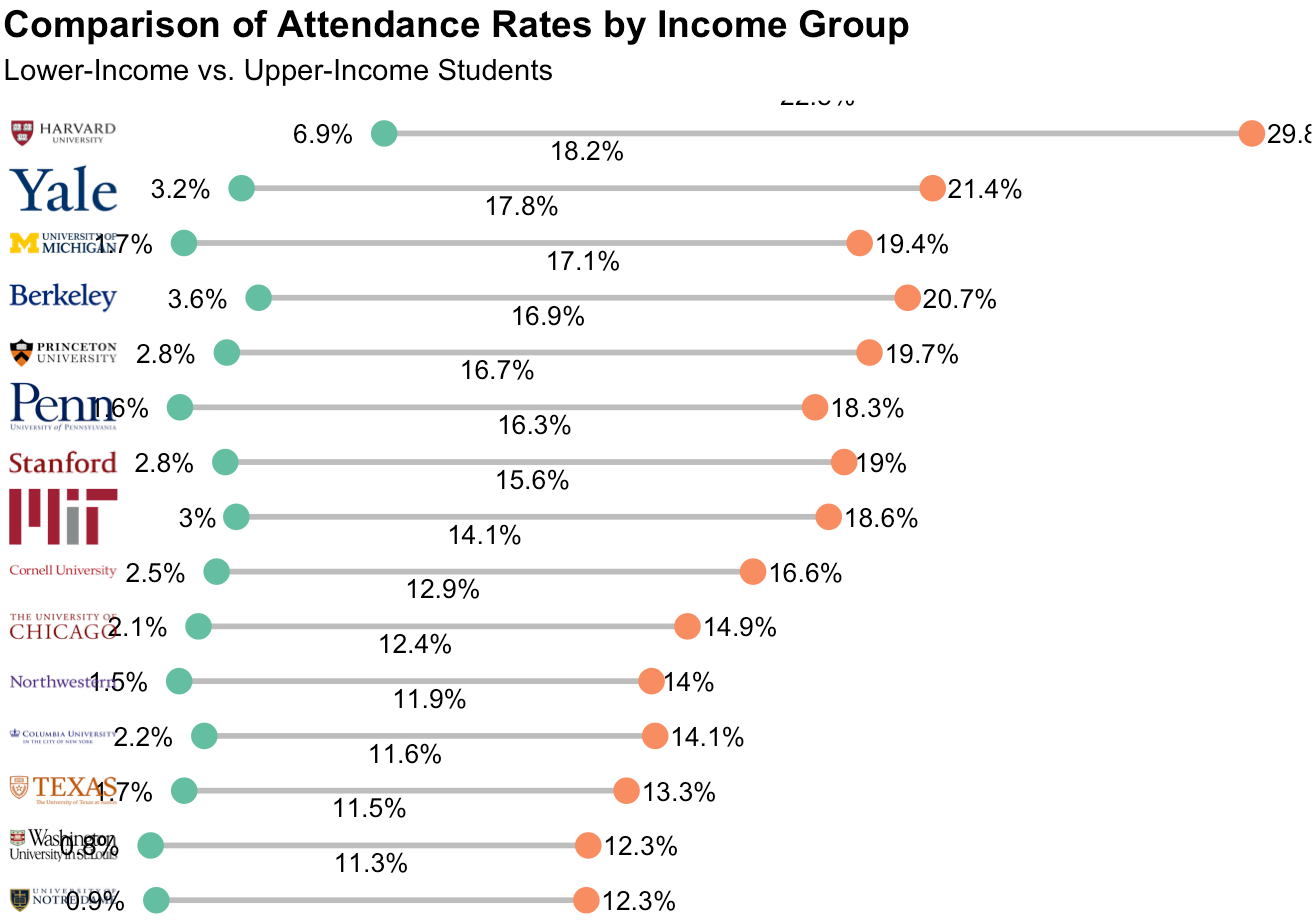
```
# Combine the two datasets: schools_for_poor_kids and schools_for_super_rich_kids
attendance_combined <- schools_for_poor_kids %>%
  rename(lower_income_attendance = weighted_attendance_rate) %>%  # Rename column for
  left_join(schools_for_super_rich_kids %>%
              rename(upper_income_attendance = weighted_attendance_rate), # Rename co
            by = "name") # Merge data frames by school name

# Calculate the gap between upper and lower-income attendance rates
attendance_combined <- attendance_combined %>%
  mutate(gap = upper_income_attendance - lower_income_attendance) # Calculate the gap

# Sort by the largest gap
attendance_combined <- attendance_combined %>%
  arrange(desc(gap)) # Arrange schools in descending order of the gap
attendance_combined
```

name <chr>
Harvard University
Yale University
University Of Michigan - Ann Arbor
University Of California, Berkeley
Princeton University
University Of Pennsylvania
Stanford University
Massachusetts Institute Of Technology
Cornell University
University Of Chicago

1-10 of 139 rows | 1-1 of 4 columns Previous **1** 2 3 4 5 6 ... 14 Next

```r
# Filter the top 15 universities by gap
attendance_combined_top <- attendance_combined %>%
  top_n(15, wt = gap) %>% # Select the top 15 schools with the largest gaps
  mutate(lower_income_attendance_percent = lower_income_attendance * 100,
         upper_income_attendance_percent = upper_income_attendance * 100) %>%
  arrange(desc(gap)) # Arrange by gap

# Get images for graph by creating a dataframe with university names and logo URLs
logos <- data.frame(
  name = c("Harvard University", "Yale University", "University Of Michigan - Ann Arb
           "University Of California, Berkeley", "Princeton University", "University
           "Stanford University", "Massachusetts Institute Of Technology",
           "Cornell University", "University Of Chicago", "Northwestern University",
           "Columbia University In The City Of New York", "University Of Texas At Aus
           "Washington University In St. Louis", "University Of Notre Dame"),
 logo_url2 = c("/Tidy Tuesday/2024-09-10

)

# Combine the attendance data with the logos
attendance_combined_top <- attendance_combined_top %>%
  left_join(logos, by = "name") # Join the logos dataframe with the attendance data

# Create a Dumbbell Chart for the top 15 universities
ggplot(attendance_combined_top, aes(x = lower_income_attendance_percent,
                                    xend = upper_income_attendance_percent,
                                    y = reorder(name, gap))) +
  geom_segment(aes(x = lower_income_attendance_percent, xend = upper_income_attendanc
                   y = reorder(name, gap), yend = reorder(name, gap)),
               color = "gray", size = 1.0) + # Line connecting the points
  geom_point(aes(x = lower_income_attendance_percent),
             color = brewer.pal(3, "Set2")[1], size = 4) + # Lower-income point
  geom_point(aes(x = upper_income_attendance_percent),
             color = brewer.pal(3, "Set2")[2], size = 4) + # Upper-income point
  geom_text(aes(x = lower_income_attendance_percent,
                label = paste0(round(lower_income_attendance_percent, 1), "%")), hjus
  geom_text(aes(x = upper_income_attendance_percent, label = paste0(round(upper_incom
  geom_text(aes(label = paste0(round(gap * 100, 1), "%"), x = (lower_income_attendanc
  geom_image(aes(x = -1.5, y = reorder(name, gap), image = logo_url2), size = 0.13, h
  labs(title = "Comparison of Attendance Rates by Income Group",
       subtitle = "Lower-Income vs. Upper-Income Students",
```

```
       x = "Attendance Rate (%)",
       y = NULL) + # Remove the y-axis title
  theme_minimal() +
  theme(axis.text.x = element_blank(), # Remove x-axis text
        axis.text.y = element_blank(), # Remove y-axis text
        axis.title.x = element_blank(),
        plot.title = element_text(size = 14, face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



We see that the gap between upper-income and lower-income student attendance rates varies across the top 15 universities. The data shows a significant economic divide in attendance rates at prestigious universities like Harvard, Yale, and Princeton, which have the largest gaps, showing a significant difference in attendance rates between upper-income and lower-income students. The larger the gap, the greater the disparity in attendance rates between upper-income and lower-income students. This disparity indicates that while these universities are accessible to students from various economic backgrounds, students from higher-income households are more likely to attend.

Plot 2 Mean Attendance Gap by Institution Tier

Compare the attendance gaps with similar institutions to understand whether certain universities are outliers or if similar patterns exist across peers.

```
# Merge CA dataset with attendance_combined on 'name'
merged_ca <- merge(attendance_combined, ca, by = "name", all.x = TRUE)

# Filter schools from previous plot
```

```r
my_15 <- merged_ca %>%
  filter(name %in% c("Harvard University", "Yale University", "University Of Michigar
  select(name, tier, lower_income_attendance, upper_income_attendance, gap)
unique(my_15)
```

	name
	<chr>
1	Columbia University In The City Of New York
15	Cornell University
29	Harvard University
43	Massachusetts Institute Of Technology
57	Northwestern University
71	Princeton University
85	Stanford University
99	University Of California, Berkeley
113	University Of Chicago
127	University Of Michigan - Ann Arbor

1-10 of 15 rows | 1-2 of 6 columns

Previous 1 2 Next

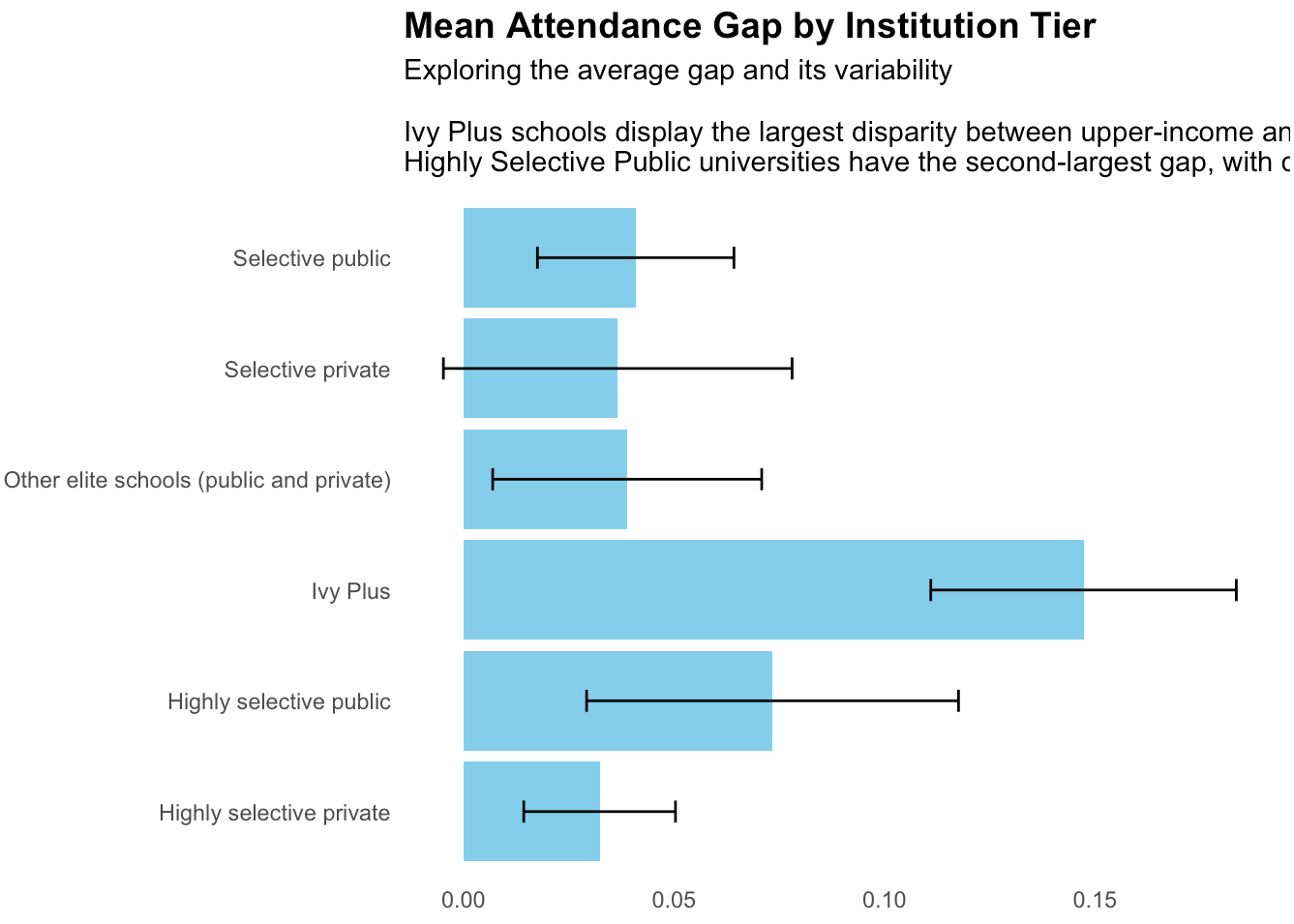
```r
# Summary statistics of the 'gap' variable, grouped by school tier
tier_summary <- merged_ca %>%
  group_by(tier) %>% # Group data by 'tier'
  summarise(
    min_gap = min(gap, na.rm = TRUE), # Minimum gap
    mean_gap = mean(gap, na.rm = TRUE), # Mean gap
    median_gap = median(gap, na.rm = TRUE), # Median gap
    sd_gap = sd(gap, na.rm = TRUE), # SD gap
    max_gap = max(gap, na.rm = TRUE)) # Max gap
tier_summary
```

tier	min_gap	mean_gap
<chr>	<dbl>	<dbl>
Highly selective private	0.005193667	0.03232630
Highly selective public	0.015398666	0.07339274
Ivy Plus	0.094463045	0.14724201
Other elite schools (public and private)	0.005593857	0.03887455
Selective private	0.006623512	0.03663165
Selective public	0.007736997	0.04089446

6 rows | 1-3 of 6 columns

```r
# Box plot of gap by tier
# ggplot(merged_ca, aes(x = tier, y = gap)) +
#   geom_boxplot() +
#   labs(x = "Selectivity Tier", y = "Attendance Gap") +
#   coord_flip()

# Scatter plot of lower vs. upper income attendance by tier
```

```r
# ggplot(merged_ca, aes(x = lower_income_attendance, y = upper_income_attendance, col
#   geom_point() +
#   labs(x = "Lower Income Attendance", y = "Upper Income Attendance", color = "Selec


# Create the bar plot mean attendance gap by tier
ggplot(tier_summary, aes(x = tier, y = mean_gap)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_errorbar(aes(ymin = mean_gap - sd_gap, ymax = mean_gap + sd_gap), width = 0.2)
  labs(title = "Mean Attendance Gap by Institution Tier",
       subtitle = "Exploring the average gap and its variability\n\nIvy Plus schools

       x = NULL,
       y = "Mean Gap") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        plot.title = element_text(size = 14, face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  coord_flip()
```




```r
# Selective Public Institutions show relatively small gaps with some variability in e

# Selective Private show relatively small gaps, with some schools showing very equita

# Other Elite Schools (Public and Private): These schools display moderate gaps with
```

```
# Ivy Plus Institutions: These institutions display the largest disparity between upp

# Highly Selective Public Institutions: These universities have the second-largest ga

# Highly Selective Private Institutions: These institutions show a more balanced dist
```

“Ivy Plus” schools, such as Harvard, Yale, Cornell, Columbia, Princeton, Stanford, University of Chicago, University of Pennsylvania, and MIT, display the largest disparity in attendance between upper- and lower-income students, with a mean gap of 0.147. This gap is significantly larger than that of other institution types, and the maximum gap of 0.228 further highlights the unequal access to these prestigious schools. Even the most equitable “Ivy Plus” schools show a minimum gap of 0.0945, indicating that income disparity remains an issue even at the best-performing schools in this category.

“Highly selective public” universities like the University of California, Berkeley, the University of Michigan, and the University of Texas at Austin exhibit the second-largest gap in attendance rates between income groups, with a mean gap of 0.0734. Although this disparity is smaller than that seen at “Ivy Plus” institutions, the relatively high standard deviation of 0.0442 points to significant variation in equity among these schools, with some being much more accessible to lower-income students than others.

In contrast, “highly selective private” institutions like Gonzaga, Syracuse, Baylor, and TCU show a much smaller mean gap of 0.0323, suggesting a more balanced distribution of students from different income backgrounds. The low variability within this category, with a standard deviation of just 0.0180, further indicates that these schools tend to be more consistent in their access.

“Selective private” (BYU, Howard, and Loyola Marymount) and “selective public” schools (Auburn, Florida State, and Alabama), with mean gaps of 0.0366 and 0.0409, respectively, reveal only slightly higher disparities than “highly selective private” schools. These institutions exhibit relatively small gaps, and their low minimum gaps suggest that some schools in these categories are very equitable in attendance for lower-income students.

“Other elite schools” like Notre Dame, Washington University, and Northwestern University display a moderate mean gap of 0.0389, higher than “highly selective private schools” but still relatively small. However, the higher standard deviation of 0.0319 indicates greater variability in income-based attendance equity within this category.

Overall, lower-income students remain vastly underrepresented across all types of institutions. For instance, “Ivy Plus” schools have an average lower-income attendance rate of only 0.0252, while “highly selective public” institutions show an even lower rate of 0.0124. Conversely, upper-income students are overrepresented, with mean attendance rates of 0.172 at “Ivy Plus” schools and 0.0858 at “highly selective public” schools.

Plot 3 Application vs. Attendance Rates

```
# Define lower-income and upper-income groups
lower_income_bins <- c(0, 10, 20, 30)
upper_income_bins <- c(80, 90, 95, 96, 97, 98, 99, 99.9, 100)

# rel_apply: How many students with similar test scores apply to a college compared t
# rel_attend: How many students with similar test scores attend a college compared to
# rel_apply_instate: How many in-state students apply to a public college compared to
# rel_attend_instate: How many in-state students attend a public college compared to
# rel_apply_oostate: How many out-of-state students apply to a public college compare
```

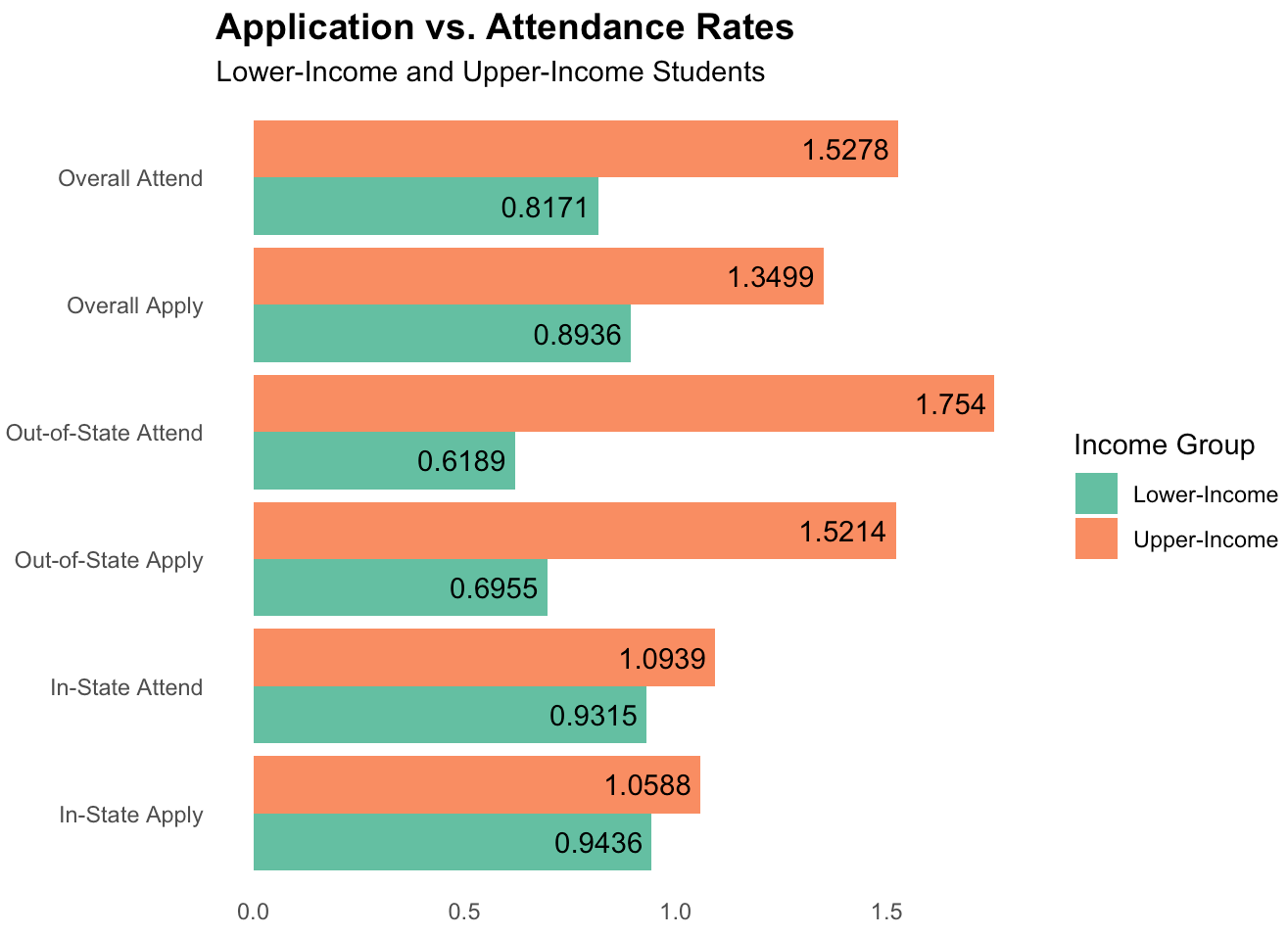
```r
# rel_attend_oostate: How many out-of-state students attend a public college compared

# Calculate mean application and attendance rates for lower-income students
lower_income_analysis <- merged_ca %>%
  filter(par_income_bin <= 30) %>% # Filter data for lower-incom
  summarize(
    mean_rel_apply = mean(rel_apply, na.rm = TRUE),
    mean_rel_attend = mean(rel_attend, na.rm = TRUE),
    mean_rel_apply_instate = mean(rel_apply_instate, na.rm = TRUE),
    mean_rel_attend_instate = mean(rel_attend_instate, na.rm = TRUE),
    mean_rel_apply_oostate = mean(rel_apply_oostate, na.rm = TRUE),
    mean_rel_attend_oostate = mean(rel_attend_oostate, na.rm = TRUE)
  )

# Calculate mean application and attendance rates for upper-income students
upper_income_analysis <- merged_ca %>%
  filter(par_income_bin >= 80) %>% # Filter data for upper-income
  summarize(
    mean_rel_apply = mean(rel_apply, na.rm = TRUE),
    mean_rel_attend = mean(rel_attend, na.rm = TRUE),
    mean_rel_apply_instate = mean(rel_apply_instate, na.rm = TRUE),
    mean_rel_attend_instate = mean(rel_attend_instate, na.rm = TRUE),
    mean_rel_apply_oostate = mean(rel_apply_oostate, na.rm = TRUE),
    mean_rel_attend_oostate = mean(rel_attend_oostate, na.rm = TRUE)
  )

# Combine the results
comparison <- bind_rows(
  lower_income_analysis %>% mutate(income_group = "Lower-Income"),
  upper_income_analysis %>% mutate(income_group = "Upper-Income")
) %>%
  select(income_group, everything()) # Reorganize columns
comparison
```

income_group	mean_rel_apply	mean_rel_attend	mean_rel_apply_instate
<chr>	<dbl>	<dbl>	<dbl>
Lower-Income	0.8935817	0.8171062	0.9435895
Upper-Income	1.3498875	1.5278338	1.0587926

2 rows | 1-4 of 7 columns

```r
# Create a DataFrame with the summary data
data <- data.frame(
  Income_Group = rep(c("Lower-Income", "Upper-Income"), each = 6),
  Category = rep(c("Overall Apply", "Overall Attend", "In-State Apply", "In-State Att
  Rate = c(0.8936, 0.8171, 0.9436, 0.9315, 0.6955, 0.6189,
           1.3499, 1.5278, 1.0588, 1.0939, 1.5214, 1.7540))

# Plot the data
ggplot(data, aes(x = Category, y = Rate, fill = Income_Group)) +
  geom_bar(stat = "identity", position = "dodge") +
    geom_text(aes(label = round(Rate, 4)), position = position_dodge(width = 0.9),
      vjust = 0.5, hjust = 1.1) +
```

```
  scale_fill_manual(values = brewer.pal(10, "Set2")) +
  labs(title = "Application vs. Attendance Rates",
       subtitle = "Lower-Income and Upper-Income Students",
       x = NULL,
       fill = "Income Group") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),  # Remove x-axis text
        #axis.text.y = element_blank(), # Remove y-axis text
        plot.title = element_text(size = 14, face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  coord_flip()
```




```
# The comparison
# Overall Mean Application Rate: 0.8936
# Overall Mean Application Rate: 1.3499

# Overall Mean Attendance Rate: 0.8171
# Overall Mean Attendance Rate: 1.5278

# In-State Mean Application Rate: 0.9436
# In-State Mean Application Rate: 1.0588

# In-State Mean Attendance Rate: 0.9315
# In-State Mean Attendance Rate: 1.0939

# Out-of-State Mean Application Rate: 0.6955
```

```
# Out-of-State Mean Application Rate: 1.5214

# Out-of-State Mean Attendance Rate: 0.6189
# Out-of-State Mean Attendance Rate: 1.7540
```

The overall mean application rate for lower-income students is 0.8936, while their mean attendance rate is 0.8171. This indicates a noticeable drop-off between application and attendance, particularly evident for out-of-state institutions. Specifically, the application rate for out-of-state colleges is 0.6955, which is higher than the attendance rate of 0.6189. This suggests that lower-income students face significant challenges in following through with attendance, possibly due to financial constraints or logistical difficulties. In contrast, the application rate for in-state colleges is slightly higher at 0.9436, with an attendance rate of 0.9315. While most in-state applicants attend, there is still a small gap between application and attendance, highlighting that barriers remain for lower-income students even with better access.

The overall mean application rate for upper-income students is 1.3499, and their mean attendance rate is even higher at 1.5278. This substantial difference shows that upper-income students are more likely to both apply and attend the colleges they apply to. The data reveals a strong follow-through on applications, with higher attendance rates than application rates. When focusing on in-state colleges, the application rate is 1.0588, which is lower than the attendance rate of 1.0939, indicating a high likelihood of attending once they apply. For out-of-state colleges, the application rate is 1.5214, with an attendance rate of 1.7540. Compared to the application rate, this substantial attendance rate suggests that upper-income students who apply to out-of-state colleges are very likely to attend. This indicates a strong follow-through and fewer barriers for upper-income students, reflecting their greater resources and access to educational opportunities.