Final Project - Data 606

Heleine Fouda

2023-12-04

The Research question:

Does parental income determines admission and attendance chances into highly selective U.S. colleges? Controlling for SAT/ACT scores, this analysis will examine that question at the application, admission and attendance phases of colleges enrollment.The analysis will also explore the specific probability of a student getting into a highly selective U.S. colleges given a parental income range between 70-80.

The data

The data sets used in this project are from a 2023 Harvard university study conducted by Raj Chetty, David J. Deming, and John N. Friedman.

The data were collected from 1996 to 2021 and combine anonymized admissions data from several private and public colleges linked to parents' and students' income tax records and students' SAT/ACT scores. The statistics reported in the original study are from 139 colleges in the U.S. comprised of the following groups: Ivy- Plus colleges(the eight Ivy leagues colleges + Stanford, Duke, MIT and Chicago), as well as highly selective private colleges and flagship public colleges. The primary measure of parental income is total household-level pre-tax income. The initial study also constructed three distinct measures of attendance, application and attendance conditional on application. The cases in the primary data set are U.S. selective colleges and universities and there are a total of 1946 cases or observations in it. Our analysis uses the following categorical variable as our response variable: parent Income Percentile; and the following numeric variables as our explanatory variables: rel_apply, attend, and rel_attend.

Getting Started: Loading Libraries

Importing and preparing the data

Let's first take a peek at the distribution of test scores conditional of parent income.

The table below reveals that there are very few students from low- income families with sufficient high - test scores to get into highly selective colleges in the first place.

```
test_scores <- read_csv("https://raw.githubusercontent.com/Heleinef/Data-Science-Master_Heleine/main/Di
spec(test_scores)
```

```
## cols(
##   `Parent Income Percenticle:` = col_character(),
##   `0-10` = col_character(),
##   `0-20` = col_character(),
##   `20-30` = col_character(),
##   `30-40` = col_character(),
##   `40-50` = col_character(),
##   `50-60` = col_character(),
##   `60-70` = col_character(),
##   `70-80` = col_character(),
##   `80-90` = col_character(),
##   `90-95` = col_character(),
##   `95-96` = col_character(),
##   `96-97` = col_character(),
##   `97-98` = col_character(),
##   `98-99` = col_character(),
##   `99-99.9` = col_character(),
##   `Top 0.1%` = col_character(),
##   `Total%` = col_character(),
##   ...19 = col_logical()
## )
```

```r
tibble(test_scores)
```

```
## # A tibble: 13 x 19
##    Parent Income Percent~1 `0-10` `0-20` `20-30` `30-40` `40-50` `50-60` `60-70`
##    <chr>                   <chr>  <chr>  <chr>   <chr>   <chr>   <chr>   <chr>
##  1 Distribution of Test S~ <NA>   <NA>   <NA>    <NA>    <NA>    <NA>    <NA>
##  2 1500-1600 (or ACT of 3~ 0.0%   0.0%   0.1%    0.1%    0.1%    0.2%    0.3%
##  3 1400-1490 (or ACT of 3~ 0.1%   0.1%   0.2%    0.2%    0.3%    0.5%    0.8%
##  4 1300-1390 (or ACT of 2~ 0.4%   0.4%   0.6%    0.7%    1.1%    1.7%    2.4%
##  5 1200-1290 (or ACT of 2~ 0.7%   0.8%   1.1%    1.4%    2.0%    2.9%    4.0%
##  6 1100-1190 (or ACT of 2~ 1.6%   2.0%   2.5%    3.1%    4.3%    6.0%    8.1%
##  7 1000-1090 (or ACT of 2~ 2.3%   2.9%   3.5%    4.3%    5.7%    7.4%    9.4%
##  8 900-990 (or ACT of 19 ~ 3.8%   5.2%   6.1%    7.2%    8.7%    10.4%   12.3%
##  9 800-890 (or ACT of 17 ~ 3.9%   5.6%   6.4%    7.1%    7.8%    8.3%    8.8%
## 10 700-790 (or ACT of 15 ~ 3.6%   5.4%   5.8%    6.0%    5.9%    5.7%    5.3%
## 11 600-690 (or ACT of 13 ~ 2.5%   3.7%   3.8%    3.7%    3.3%    2.9%    2.4%
## 12 below 600 (or ACT belo~ 1.8%   1.7%   1.6%    1.4%    1.0%    0.8%    0.6%
## 13 Did not take SAT or ACT 79.8%  71.9%  68.3%   64.5%   59.5%   53.1%   45.6%
## # i abbreviated name: 1: `Parent Income Percentile:`
## # i 11 more variables: `70-80` <chr>, `80-90` <chr>, `90-95` <chr>,
## #   `95-96` <chr>, `96-97` <chr>, `97-98` <chr>, `98-99` <chr>,
## #   `99-99.9` <chr>, `Top 0.1%` <chr>, `Total%` <chr>, ...19 <lgl>
```

The above table shows that among students scoring high on the SAT/ACT, only 3.7% come from the lowest quintile of parent income.This means that there are very few students from low- income families with sufficient high - test scores to get into highly selective colleges to begin with.

Now, let's focus on colleges admissions.

```r
names(CollegesAdmissions)
```

```
##  [1] "super_opeid"                        "name"
##  [3] "par_income_bin"                     "par_income_lab"
##  [5] "attend"                             "stderr_attend"
##  [7] "attend_level"                       "attend_sat"
##  [9] "stderr_attend_sat"                  "attend_level_sat"
## [11] "rel_apply"                          "stderr_rel_apply"
## [13] "rel_attend"                         "stderr_rel_attend"
## [15] "rel_att_cond_app"                   "rel_apply_sat"
## [17] "stderr_rel_apply_sat"               "rel_attend_sat"
## [19] "stderr_rel_attend_sat"              "rel_att_cond_app_sat"
## [21] "attend_instate"                  "stderr_attend_instate"
## [23] "attend_level_instate"            "attend_instate_sat"
## [25] "stderr_attend_instate_sat"       "attend_level_instate_sat"
## [27] "attend_oostate"                  "stderr_attend_oostate"
## [29] "attend_level_oostate"            "attend_oostate_sat"
## [31] "stderr_attend_oostate_sat"       "attend_level_oostate_sat"
## [33] "rel_apply_instate"               "stderr_rel_apply_instate"
## [35] "rel_attend_instate"              "stderr_rel_attend_instate"
## [37] "rel_att_cond_app_instate"        "rel_apply_oostate"
## [39] "stderr_rel_apply_oostate"        "rel_attend_oostate"
## [41] "stderr_rel_attend_oostate"       "rel_att_cond_app_oostate"
## [43] "rel_apply_instate_sat"           "stderr_rel_apply_instate_sat"
## [45] "rel_attend_instate_sat"          "stderr_rel_attend_instate_sat"
## [47] "rel_att_cond_app_instate_sat"    "rel_apply_oostate_sat"
## [49] "stderr_rel_apply_oostate_sat"    "rel_attend_oostate_sat"
## [51] "stderr_rel_attend_oostate_sat"   "rel_att_cond_app_oostate_sat"
## [53] "attend_unwgt"                    "stderr_attend_unwgt"
## [55] "attend_unwgt_level"              "attend_unwgt_instate"
## [57] "stderr_attend_unwgt_instate"     "attend_unwgt_oostate"
## [59] "stderr_attend_unwgt_oostate"     "attend_unwgt_level_instate"
## [61] "attend_unwgt_level_oostate"      "rel_attend_unwgt"
## [63] "rel_apply_unwgt"                 "stderr_rel_attend_unwgt"
## [65] "stderr_rel_apply_unwgt"          "rel_att_cond_app_unwgt"
## [67] "rel_attend_unwgt_instate"        "rel_attend_unwgt_oostate"
## [69] "stderr_rel_attend_unwgt_instate" "stderr_rel_attend_unwgt_oostate"
## [71] "rel_apply_unwgt_instate"         "rel_apply_unwgt_oostate"
## [73] "stderr_rel_apply_unwgt_instate"  "stderr_rel_apply_unwgt_oostate"
## [75] "rel_att_cond_app_unwgt_instate"  "rel_att_cond_app_unwgt_oostate"
## [77] "public"                          "flagship"
## [79] "tier"                            "tier_name"
## [81] "test_band_tier"
```

```r
# skim main-df
skim(CollegesAdmissions)
```

Table 1: Data summary

Name	CollegesAdmissions
Number of rows	1946
Number of columns	81
Column type frequency:	
character	6
numeric	75
Group variables	None

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
public	0	1	6	7	0	2	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
tier	0	1	8	40	0	6	0
tier_name	0	1	8	40	0	6	0
test_band_tier	0	1	6	21	0	6	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
super_opeid	0	1.00	2528.51	1344.92	108.00	1536.00	2536.00	3223.00	11649.00	
par_income_bin	0	1.00	78.17	28.04	10.00	65.00	94.00	98.50	100.00	
attend	2	1.00	0.00	0.01	0.00	0.00	0.00	0.01	0.05	
stderr_attend	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	
attend_level	0	1.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	
attend_sat	294	0.85	0.00	0.00	0.00	0.00	0.00	0.01	0.04	
stderr_attend_sat	278	0.86	0.00	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	0.02	
rel_apply	0	1.00	1.17	0.56	0.07	0.83	1.05	1.34	5.87	
stderr_rel_apply	0	1.00	0.04	0.04	0.01	0.02	0.03	0.05	0.53	
rel_attend	2	1.00	1.27	0.93	0.01	0.76	1.03	1.45	10.26	
stderr_rel_attend	0	1.00	0.11	0.12	0.00	0.05	0.08	0.13	1.57	
rel_att_cond_app	2	1.00	1.03	0.27	0.05	0.88	1.00	1.15	3.06	
rel_apply_sat	278	0.86	1.15	0.51	0.19	0.79	1.06	1.36	4.70	
stderr_rel_apply_sat	278	0.86	0.07	0.05	0.02	0.03	0.06	0.09	0.56	
rel_attend_sat	294	0.85	1.19	0.75	0.02	0.71	0.99	1.48	8.03	
stderr_rel_attend_sat	278	0.86	0.18	0.13	0.00	0.08	0.15	0.24	0.95	
rel_att_cond_app_sat	294	0.85	1.01	0.35	0.04	0.82	0.98	1.17	3.85	
attend_instate	1334	0.31	0.13	0.08	0.00	0.06	0.13	0.18	0.41	
stderr_attend_instate	1334	0.31	0.01	0.01	0.00	0.00	0.01	0.01	0.08	
attend_level_instate	1232	0.37	0.12	0.07	0.02	0.05	0.13	0.17	0.36	
attend_instate_sat	1334	0.31	0.13	0.10	0.00	0.04	0.11	0.18	0.66	
stderr_attend_instate_sat	1334	0.31	0.02	0.02	0.00	0.00	0.01	0.02	0.15	
attend_level_instate_sat	1232	0.37	0.12	0.08	0.01	0.03	0.11	0.16	0.39	
attend_oostate	1336	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.02	
stderr_attend_oostate	1334	0.31	0.00	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	0.00	
attend_oostate_sat	1346	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.01	
stderr_attend_oostate_sat	1334	0.31	0.00	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	0.00	
rel_apply_instate	1334	0.31	1.01	0.17	0.27	0.91	1.00	1.12	1.70	
stderr_rel_apply_instate	1334	0.31	0.03	0.02	0.01	0.01	0.02	0.03	0.13	
rel_attend_instate	1334	0.31	1.03	0.30	0.14	0.86	1.01	1.20	2.19	
stderr_rel_attend_instate	1334	0.31	0.06	0.04	0.01	0.03	0.05	0.07	0.31	
rel_att_cond_app_instate	1334	0.31	1.01	0.19	0.37	0.90	1.00	1.11	1.70	
rel_apply_oostate	1334	0.31	1.19	0.55	0.32	0.75	1.10	1.54	4.14	
stderr_rel_apply_oostate	1334	0.31	0.05	0.04	0.01	0.02	0.03	0.06	0.26	
rel_attend_oostate	1336	0.31	1.30	0.87	0.24	0.63	1.08	1.74	5.86	
stderr_rel_attend_oostate	1334	0.31	0.14	0.11	0.00	0.06	0.11	0.18	0.75	
rel_att_cond_app_oostate	1336	0.31	1.02	0.28	0.34	0.82	1.01	1.18	2.42	
rel_apply_instate_sat	1334	0.31	1.02	0.20	0.31	0.90	1.00	1.16	1.57	
stderr_rel_apply_instate_sat	1334	0.31	0.06	0.04	0.02	0.03	0.05	0.07	0.28	
rel_attend_instate_sat	1334	0.31	1.07	0.44	0.09	0.79	1.00	1.27	2.95	
stderr_rel_attend_instate_sat	1334	0.31	0.14	0.10	0.03	0.07	0.11	0.16	0.71	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
rel_att_cond_app_instate_sat	1334	0.31	1.03	0.30	0.15	0.86	1.00	1.15	2.17	
rel_apply_oostate_sat	1334	0.31	1.30	0.70	0.28	0.73	1.12	1.73	4.21	
stderr_rel_apply_oostate_sat	1334	0.31	0.11	0.09	0.02	0.04	0.08	0.15	0.67	
rel_attend_oostate_sat	1346	0.31	1.49	1.21	0.06	0.55	1.07	2.09	9.01	
stderr_rel_attend_oostate_sat	1334	0.31	0.34	0.31	0.00	0.11	0.24	0.47	2.14	
rel_att_cond_app_oostate_sat	1346	0.31	1.05	0.42	0.10	0.76	0.99	1.25	2.59	
attend_unwgt	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	
stderr_attend_unwgt	0	1.00	0.00	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	0.00	
attend_unwgt_instate	1334	0.31	0.09	0.08	0.00	0.03	0.07	0.14	0.35	
stderr_attend_unwgt_instate	1334	0.31	0.01	0.01	0.00	0.00	0.00	0.01	0.04	
attend_unwgt_oostate	1337	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.01	
stderr_attend_unwgt_oostate	1334	0.31	0.00	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	0.24	
attend_unwgt_level_oostate	1232	0.37	0.00	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	30.80	
rel_apply_unwgt	0	1.00	1.98	1.77	0.09	0.71	1.42	2.62	13.77	
stderr_rel_attend_unwgt	0	1.00	0.20	0.35	0.00	0.03	0.07	0.22	3.21	
stderr_rel_apply_unwgt	0	1.00	0.06	0.09	0.00	0.01	0.03	0.07	1.06	
rel_att_cond_app_unwgt	1	1.00	1.06	0.39	0.01	0.84	1.00	1.20	4.95	
rel_attend_unwgt_instate	1334	0.31	1.30	0.73	0.18	0.79	1.10	1.64	4.70	
rel_attend_unwgt_oostate	1337	0.31	1.82	1.67	0.09	0.53	1.29	2.82	12.10	
stderr_rel_attend_unwgt_instate	1334	0.31	0.07	0.05	0.01	0.03	0.05	0.09	0.34	
stderr_rel_attend_unwgt_oostate	1334	0.31	0.16	0.15	0.00	0.04	0.10	0.24	0.93	
rel_apply_unwgt_instate	1334	0.31	1.19	0.43	0.35	0.86	1.10	1.49	2.40	
rel_apply_unwgt_oostate	1334	0.31	1.57	1.08	0.17	0.69	1.31	2.26	7.07	
stderr_rel_apply_unwgt_instate	1334	0.31	0.03	0.02	0.00	0.01	0.02	0.04	0.12	
stderr_rel_apply_unwgt_oostate	1334	0.31	0.05	0.04	0.00	0.01	0.03	0.07	0.31	
rel_att_cond_app_unwgt_instate	1334	0.31	1.04	0.27	0.40	0.86	1.01	1.17	2.09	
rel_att_cond_app_unwgt_oostate	1337	0.31	1.02	0.33	0.19	0.78	1.00	1.21	2.14	
flagship	0	1.00	0.21	0.41	0.00	0.00	0.00	0.00	1.00	

Summary statistics

```
# Summary - Colleges Admissions
admissions_summary <- CollegesAdmissions %>%
  group_by(par_income_lab) %>%
  summarise(across(.cols=contains("att"), list(mean, standard_deviation = sd)))
admissions_summary
```

```
## # A tibble: 14 x 109
##    par_income_lab attend_1 attend_standard_deviation stderr_attend_1
##    <chr>             <dbl>                     <dbl>           <dbl>
##  1 0-20            0.00367                   0.00464        0.000437
##  2 20-40           0.00366                   0.00441        0.000343
##  3 40-60           0.00350                   0.00366        0.000246
##  4 60-70           0.00334                   0.00321        0.000253
##  5 70-80           0.00341                   0.00318        0.000203
##  6 80-90           0.00367                   0.00335        0.000154
##  7 90-95           0.00418                   0.00367        0.000174
##  8 95-96           0.00467                   0.00407        0.000353
##  9 96-97           0.00495                   0.00424        0.000349
## 10 97-98           0.00543                   0.00466        0.000347
## 11 98-99           0.00590                   0.00493        0.000346
## 12 99-99.9         0.00668                   0.00629        0.000367
## 13 Top 0.1        NA                        NA              0.00103
## 14 Top 1           0.00679                   0.00664        0.000346
## # i 105 more variables: stderr_attend_standard_deviation <dbl>,
## #   attend_level_1 <dbl>, attend_level_standard_deviation <dbl>,
## #   attend_sat_1 <dbl>, attend_sat_standard_deviation <dbl>,
## #   stderr_attend_sat_1 <dbl>, stderr_attend_sat_standard_deviation <dbl>,
## #   attend_level_sat_1 <dbl>, attend_level_sat_standard_deviation <dbl>,
## #   rel_attend_1 <dbl>, rel_attend_standard_deviation <dbl>,
## #   stderr_rel_attend_1 <dbl>, stderr_rel_attend_standard_deviation <dbl>, ...
```

Defining elite U.S. elites colleges

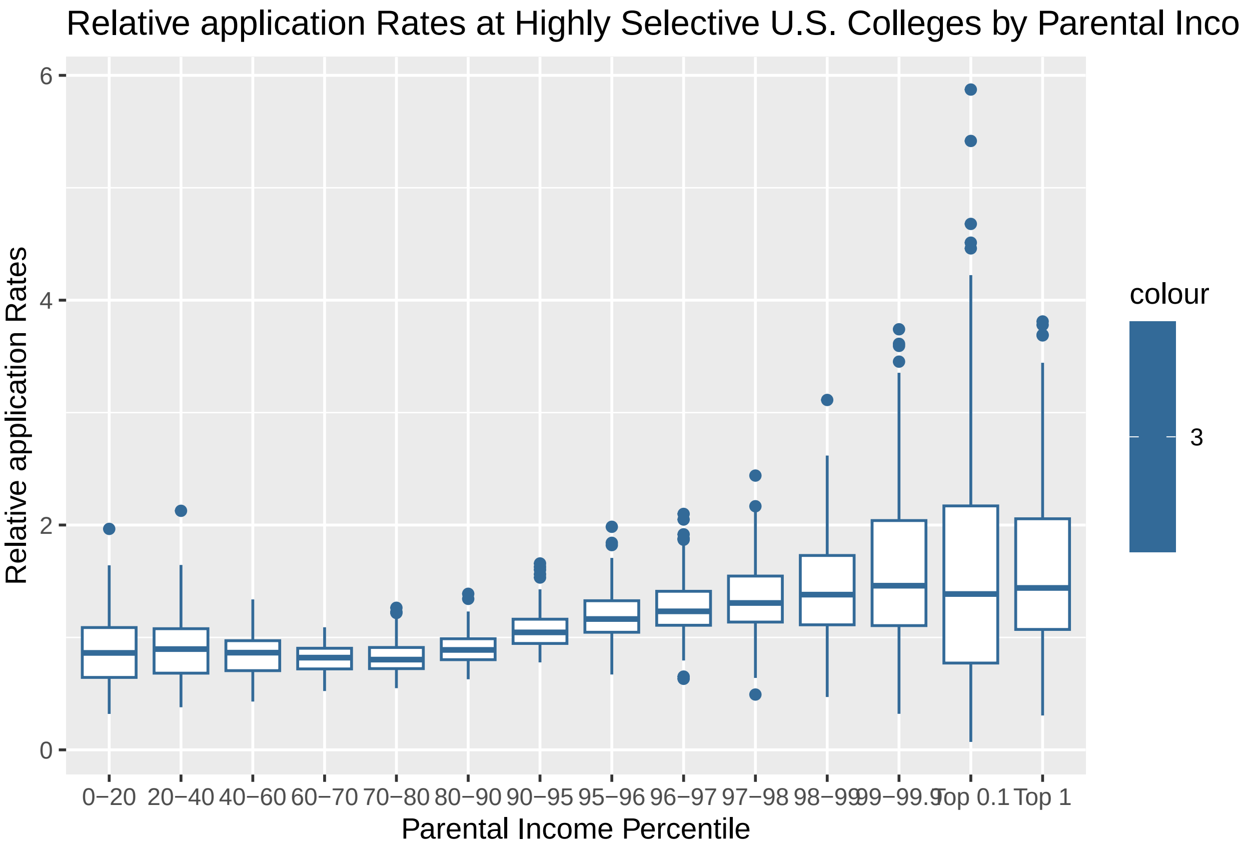
```
elite_colleges <- CollegesAdmissions %>%
  group_by(tier_name) %>%
  summarise
  elite_colleges
```

```
## # A tibble: 6 x 1
##   tier_name
##   <chr>
## 1 Highly selective private
## 2 Highly selective public
## 3 Ivy Plus
## 4 Other elite schools (public and private)
## 5 Selective private
## 6 Selective public
```

Relative application rates at Highly selective U.S. colleges by parental income

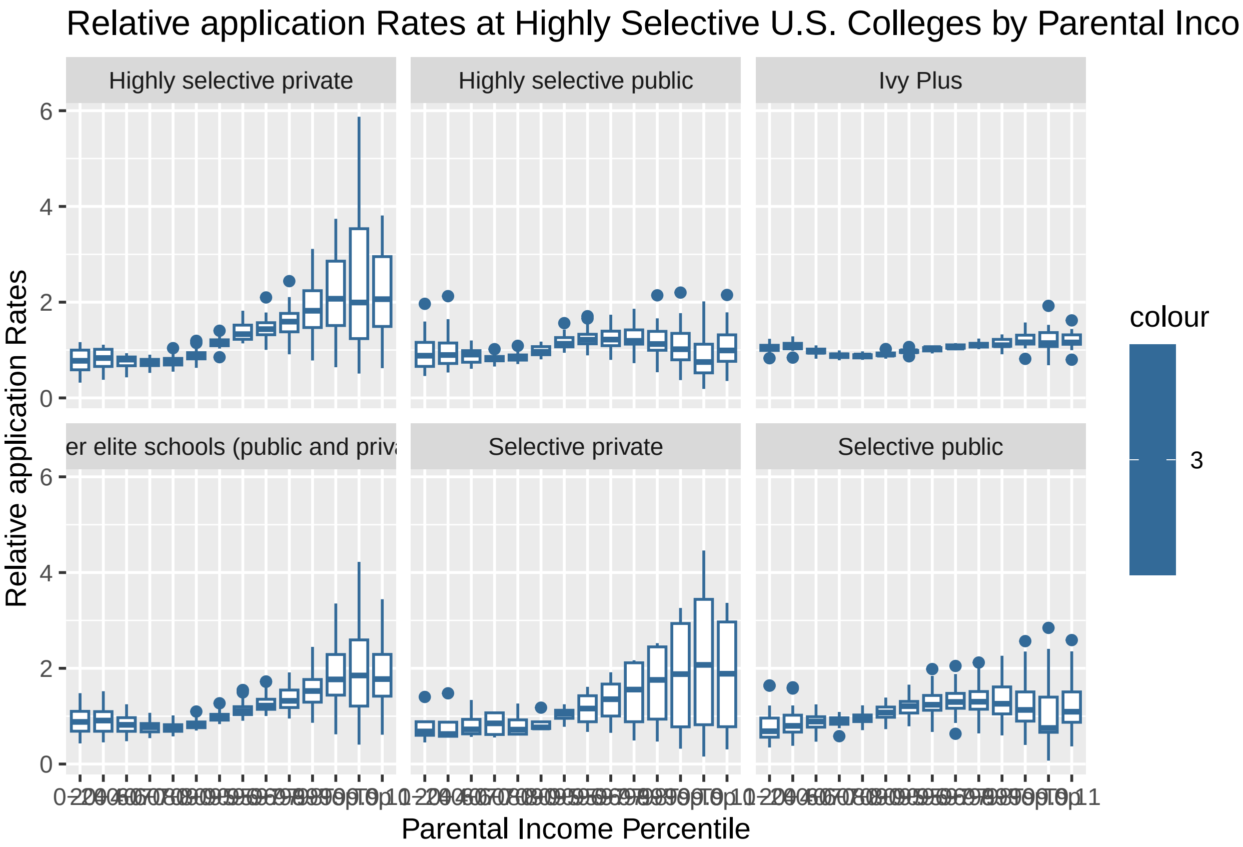
```
# Relative application rates at Highly selective U.S. colleges by parental income

ggplot(data = CollegesAdmissions, aes(x = par_income_lab, y = rel_apply, color = 3))+
  geom_boxplot()+
  labs(
    x = "Parental Income Percentile",
    y = "Relative application Rates",
    title = "Relative application Rates at Highly Selective U.S. Colleges by Parental Income",
    fill = "tier_name"
  )
```



###The boxplot above shows that, Discrepancies in admissions to elite U.S.colleges based on parental income are noticeable right from the start, at the application phase.Controling for SAT/ACT scores, the boxplot reveals that students from high parental income percentile apply to elite colleges 5 times more than do their counterparts from the low and middle parental income percentile.

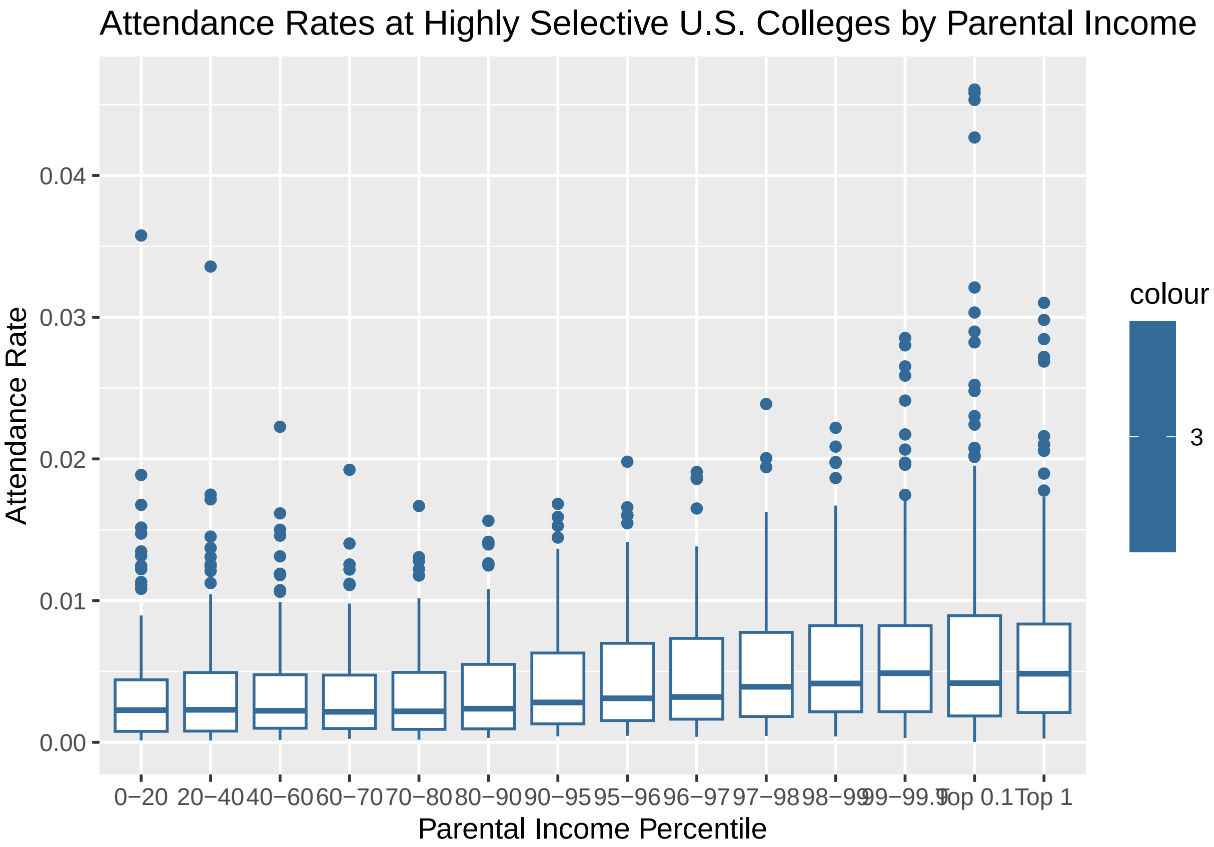
```
ggplot(data = CollegesAdmissions, aes(x = par_income_lab, y = rel_apply, color = 3))+
  geom_boxplot()+
  labs(
    x = "Parental Income Percentile",
    y = "Relative application Rates",
    title = "Relative application Rates at Highly Selective U.S. Colleges by Parental Income",
    fill = "tier_name"
  )+ facet_wrap(~tier_name)
```



Attendance rates at Highly selective colleges by parental income

```
# Attendance rates at Highly selective colleges by parental income
ggplot(data = CollegesAdmissions, aes(x = par_income_lab, y= attend, color = 3))+
  geom_boxplot()+
  labs(
    x = "Parental Income Percentile",
    y = "Attendance Rate",
    title = "Attendance Rates at Highly Selective U.S. Colleges by Parental Income",
    fill = "tier_name"
  )
```

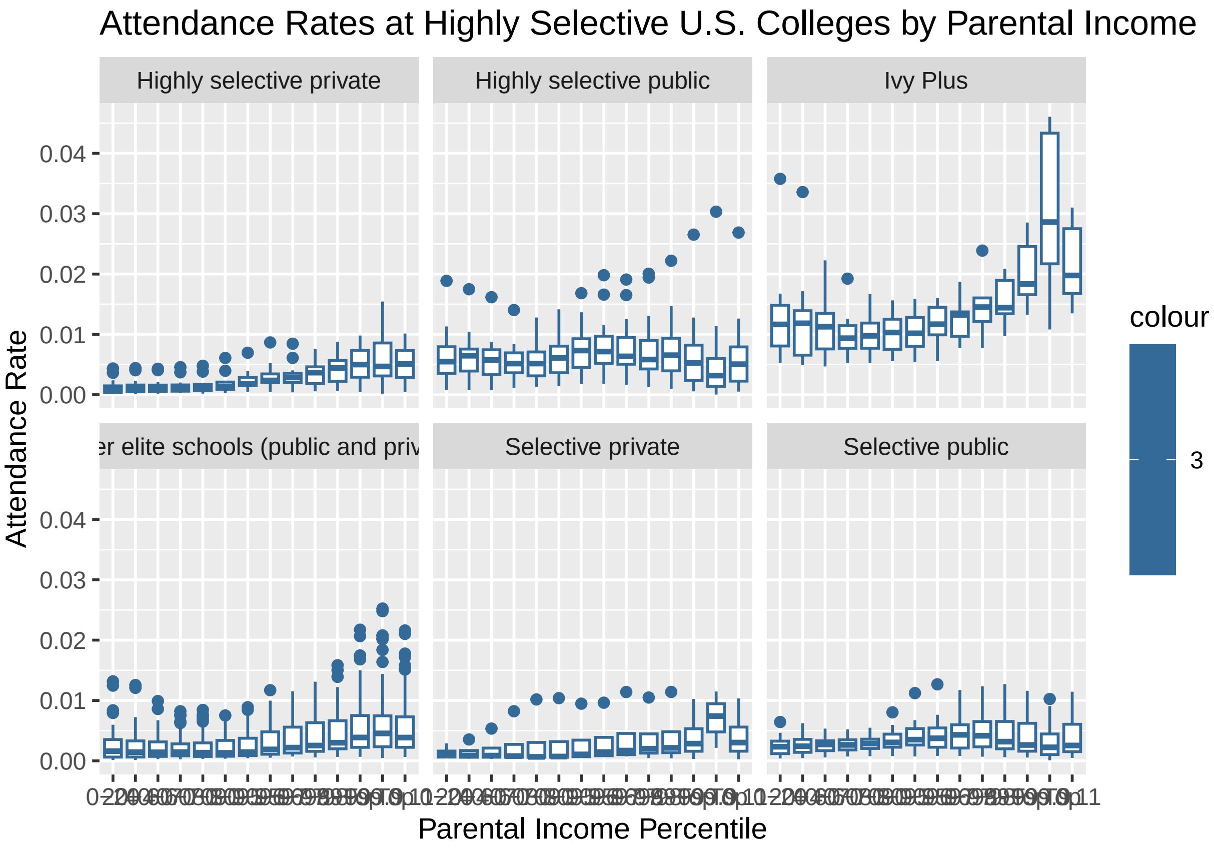
```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



The above boxplot reveals that most students from higher parental Income percentile attend highly selective U.S.colleges followed by students from very low parental income percentile. Students from middle income family count for 2% of attendance and elite U.S. colleges.

```r
# Attendance rates at Highly selective colleges by parental income
ggplot(data = CollegesAdmissions, aes(x = par_income_lab, y= attend, color = 3))+
  geom_boxplot()+
  labs(
    x = "Parental Income Percentile",
    y = "Attendance Rate",
    title = "Attendance Rates at Highly Selective U.S. Colleges by Parental Income",
    fill = "tier_name"
  )+ facet_wrap(~tier_name)
```

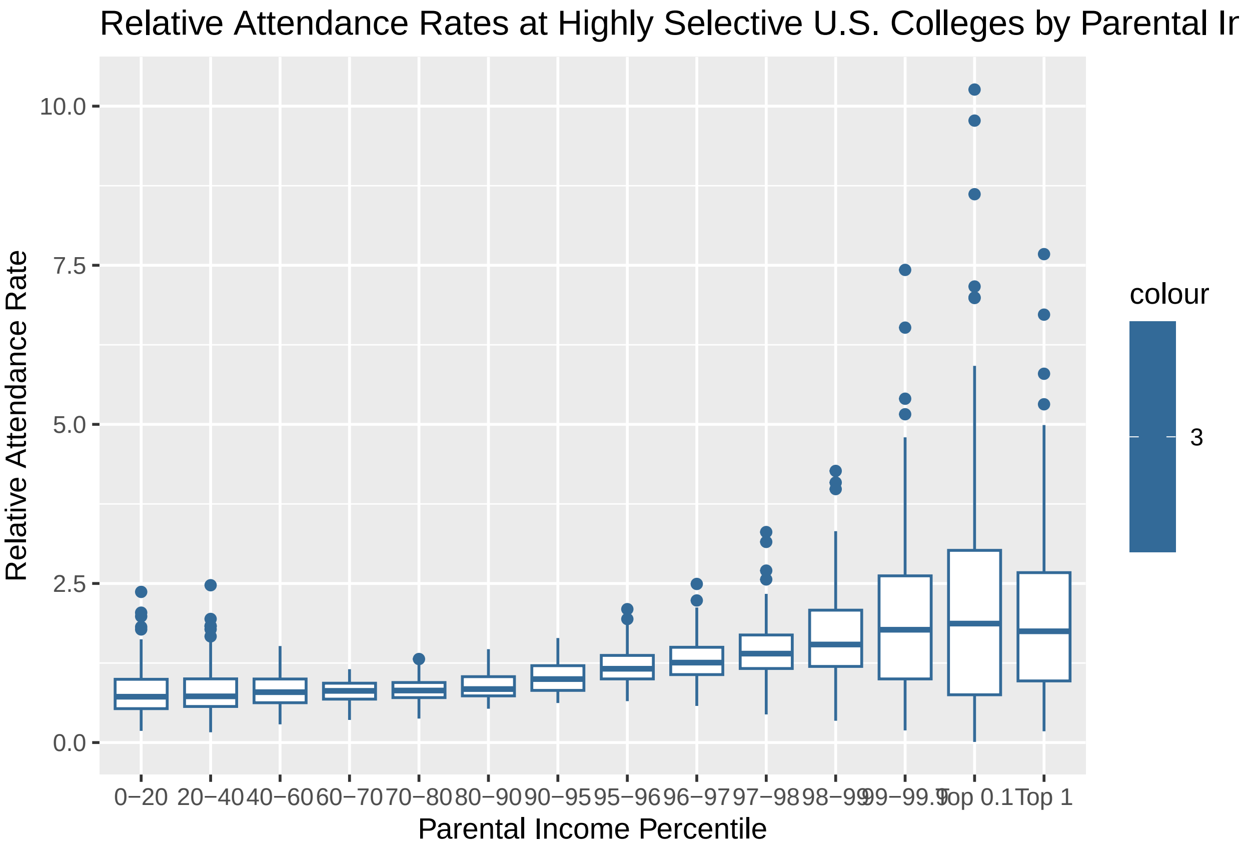
```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



Relative attendance rates at Highly selective colleges by parental income.

```r
# Relative attendance rates at Highly selective colleges by parental income
ggplot(data = CollegesAdmissions, aes(x = par_income_lab, y= rel_attend, color = 3))+
  geom_boxplot()+
  labs(
    x = "Parental Income Percentile",
    y = "Relative Attendance Rate",
    title = "Relative Attendance Rates at Highly Selective U.S. Colleges by Parental Income",
    fill = "tier_name"
  )
```

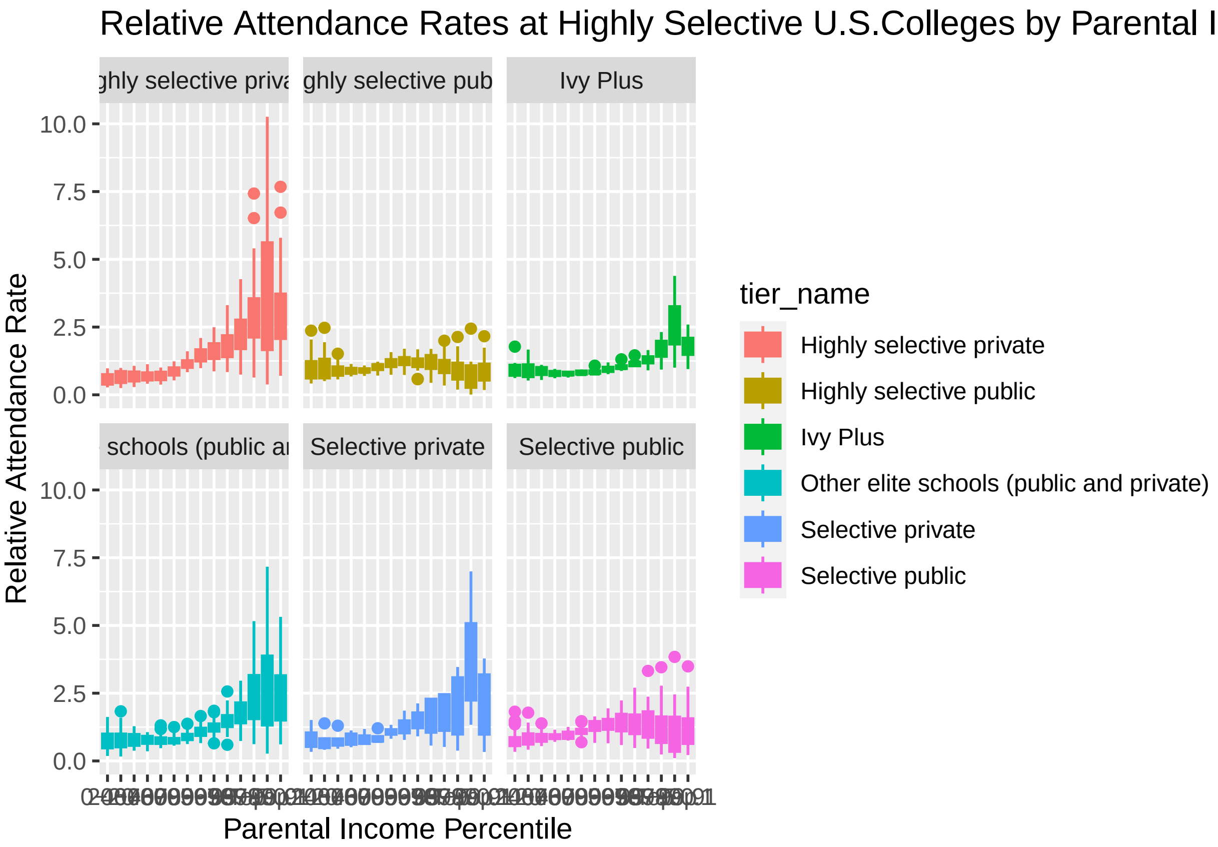
```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



Students from middle parental income ranges(40-60, 70-80, 90-95) have less than 2% chances to attend highly selective U.S. colleges, while students from top earning families have more than 5 times to get admitted than their counterparts from the middle and lower income families.

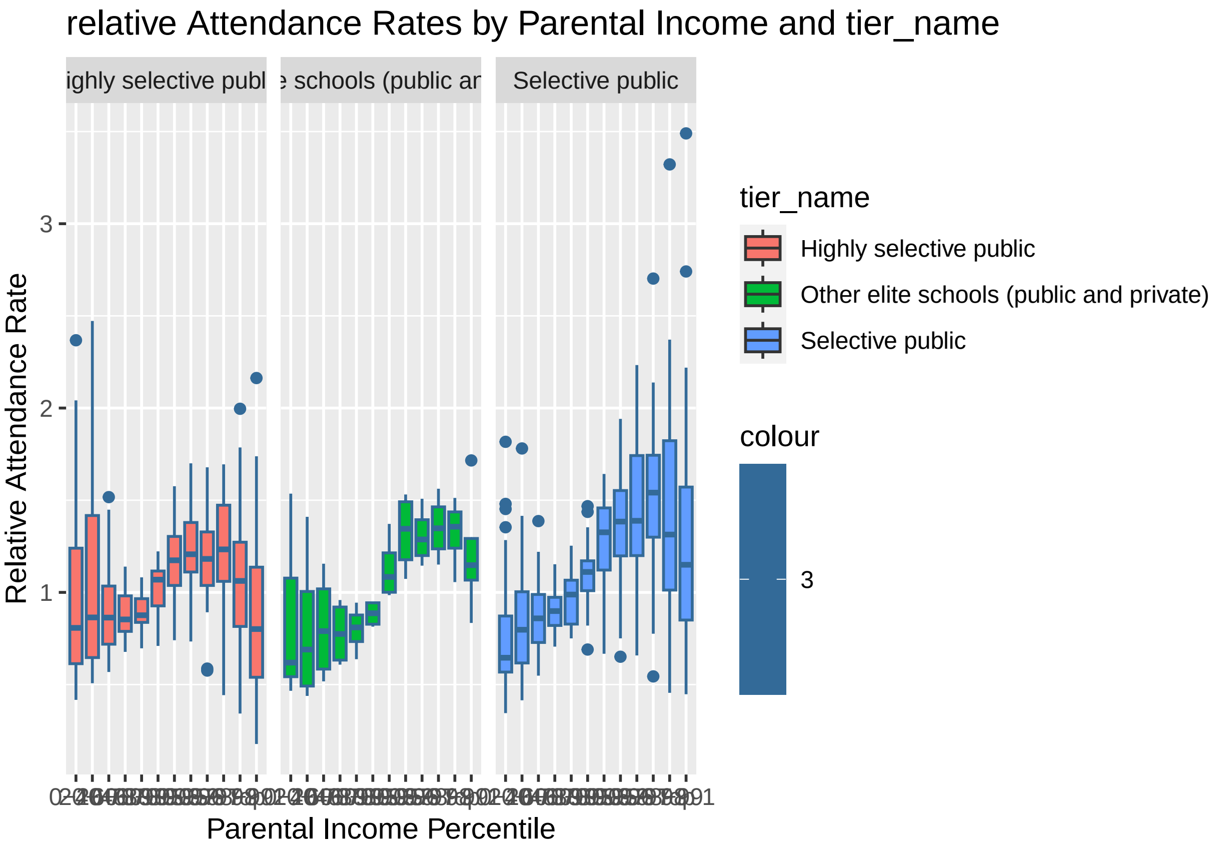
```
ggplot(CollegesAdmissions, aes(x = par_income_lab, y= rel_attend, color = tier_name, fill = tier_name))
  geom_boxplot()+
  labs(
    x = "Parental Income Percentile",
    y = "Relative Attendance Rate",
    title = "Relative Attendance Rates at Highly Selective U.S.Colleges by Parental Income",
    fill = "tier_name"
  )+ facet_wrap(~tier_name)
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



Results of relative attendance when filtering for only complete cases

```
# let's calculate relative rates by parental income based on the filtered data
ggplot(data = CollegesAdmissions_new, aes (x= par_income_lab, y = rel_attend, fill = tier_name, color =
  labs(
    x = "Parental Income Percentile",
    y = "Relative Attendance Rate",
    title = "relative Attendance Rates by Parental Income and tier_name"
  ) + facet_wrap(~tier_name)
```



Correlation tests:

Let's explore the strengths of the correlations existing between our response variable(parental income percentile) and two other variables from the Colleges Admission data set in order to understand which factor most explains difference in attendance level at selective colleges.

Positive but weak correlation between parental income and relative application to elite colleges: 0.388087

```
# Relative application to elite colleges and parental income
cor.test(CollegesAdmissions$par_income_bin,CollegesAdmissions$rel_apply)
```

```
##
##  Pearson's product-moment correlation
##
## data:  CollegesAdmissions$par_income_bin and CollegesAdmissions$rel_apply
## t = 18.566, df = 1944, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3496821 0.4251898
## sample estimates:
##       cor
## 0.388087
```

Linear regression

```
## linear modeling
CollegesAdmissions_lm1 <- lm(par_income_bin ~ rel_apply, data = CollegesAdmissions)
```

```r
summary(CollegesAdmissions_lm1)
```

```
##
## Call:
## lm(formula = par_income_bin ~ rel_apply, data = CollegesAdmissions)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83.55  -8.74  10.26  17.74  43.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   55.432      1.358   40.83   <2e-16 ***
## rel_apply     19.392      1.044   18.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.85 on 1944 degrees of freedom
## Multiple R-squared:  0.1506, Adjusted R-squared:  0.1502
## F-statistic: 344.7 on 1 and 1944 DF,  p-value: < 2.2e-16
```

There is a positive but moderate relationship between parental income and rel_att_cond_app_unwgt

```r
cor.test(CollegesAdmissions$par_income_bin,CollegesAdmissions$rel_att_cond_app_unwgt)
```

```
##
##  Pearson's product-moment correlation
##
## data:  CollegesAdmissions$par_income_bin and CollegesAdmissions$rel_att_cond_app_unwgt
## t = 21.174, df = 1943, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3961666 0.4684207
## sample estimates:
##       cor
## 0.4329889
```

Linear regression

```r
# linear modeling
CollegesAdmissions_lm2 <- lm(par_income_bin~rel_att_cond_app_unwgt, data = CollegesAdmissions)
summary(CollegesAdmissions_lm2)
```

```
##
## Call:
## lm(formula = par_income_bin ~ rel_att_cond_app_unwgt, data = CollegesAdmissions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.645  -12.019    8.387   17.607   55.015
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               44.811      1.676   26.73   <2e-16 ***
## rel_att_cond_app_unwgt    31.494      1.487   21.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.28 on 1943 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1875, Adjusted R-squared:  0.1871
## F-statistic: 448.3 on 1 and 1943 DF,  p-value: < 2.2e-16
```

Evaluating the predictive value of the linear models

```
# Predicted values (lm1)
CollegesAdmissions_lm1_predicted <- predict(CollegesAdmissions_lm1)
glimpse(CollegesAdmissions_lm1_predicted)
```

```
##  Named num [1:1946] 68.4 68.6 69.1 69.3 69.4 ...
##  - attr(*, "names")= chr [1:1946] "1" "2" "3" "4" ...
```

```
# Predicted values (lm2)
CollegesAdmissions_lm2_predicted <- predict(CollegesAdmissions_lm2)
glimpse(CollegesAdmissions_lm2_predicted)
```

```
##  Named num [1:1945] 81.4 75.8 82.1 85.2 85.4 ...
##  - attr(*, "names")= chr [1:1945] "1" "2" "3" "4" ...
```

Evaluating the probability of getting accepted into a U.S.elite college with a middle class parental income range between 70-80

The Uniform Probability Distribution: if we assume all applicants with the same SAT/ACT scores have an equal chance of being admitted into elite U.S.colleges regardless of their parental income, Then the probability of a student from a middle class parental income between 70-80 being admitted is the same as the probability of a student from the top 1% and can be written as follow:

P(applicant) = 1 / n where n is the total number of applicants 1.

```
# Assuming a class with n =100
set.seed(337)
n <- 100
applicant <- 70-80

1/n
```

```
## [1] 0.01
```

Conditional Probability: non - academic requirements at elites significantly reduces the probability that a student from a middle income range will get into an elite U.S. college

Given that elite U.S. colleges have different admission criteria beyond academic performance, such as extracurricular activities and/or legacy preferences. The actual probability of admission may vary depending on these criteria.Therefore the probability of admission for a student from a middle class parental income between 70-80 being admitted in an elite college could be read as followed:

P(A | B) = P(A and B) / P(B)

where A represents the event of an applicant being admitted, and B represents the event of an applicant meeting the admission criteria. Therefore, the probability of an applicant being admitted given that they meet the admission criteria is given by:

P(admitted | meets criteria) = P(admitted and meets criteria) / P(meets criteria)

where P(admitted | meets criteria) is the probability of an applicant being admitted given that they meet the admission criteria, P(admitted and meets criteria) is the probability of an applicant being admitted and meeting the admission criteria, and P(meets criteria) is the probability of an applicant meeting the admission criteria.

Main Findings:

1.We found disparities in application, admission and attendance rates at american elite colleges

2.Among students scoring high on the SAT/ACT, only 3.7% come from the lowest quintile of parent income.

3.From the factors we measured, we found a positive but moderate relationship between parental income and rel_att_cond_app_unwgt(0.4329889). We also found a positive but weak correlation between parental income and relative application to elite colleges: 0.388087

4.Students from middle parental income ranges(40-60, 70-80, 90-95) have less than 2% chances to attend highly selective U.S. colleges

5.There are very few students from low - income families with sufficient high - test scores to get into highly selective colleges.

6.There is a substantial under matching of middle – income & low- income students at Ivy-Plus schools.

7.Ivy-Plus colleges are more than twice as likely to admit a student from a high – income family as compared to low or middle – income families with comparable SAT/ACT scores.

8.Controlling for SAT/ACT scores, the probability of attending an elite private college is 77 times higher for children in the top 1% compared to students fromthe bottom 20%