Read in Data

- Map column values according to data dictionary
- Clean column names

```
library(plyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()
```

```
library(ggplot2)
library(gtools)

raw_df <- read.csv("../data/SY18-19_all_variables_Data.csv", check.names = FALSE)
df <- raw_df
value_mapping <- read.csv("../data/SY18-19_all_variables_ValueLabels.csv")

# convert the sector into the actual names
for (i in unique(value_mapping$VariableName)){
  temp_map_val <- value_mapping %>% subset(VariableName == i)
  df[[i]] <- mapvalues(df[[i]], from = temp_map_val$Value, to = temp_map_val$ValueLabel)
}

# clean up column names
colnames(df) <- gsub(" - ", "_", tolower(colnames(df)))
colnames(df) <- gsub("-|[[:space:]]+|/", "_", colnames(df))
colnames(df) <- gsub("\\(|\\)|\\:", "", colnames(df))
colnames(df) <- gsub("\\_(sfa1819|2018_19|drvgr2018|drvef2018|hd2018|2018_ef2018d|f1819|al2019|gasb_drv
colnames(df) <- gsub("graduation_rate_bachelor_degree_within_6_years", "gradrate_ba_6yrs", colnames(df))

# combine the financial columns which were broken up for public and private schools into one
dupe_vals <- grep('value_of_endowment_assets_at_the_beginning_of_the_fiscal_year', colnames(df))
colnames(df)[dupe_vals[1]] <- 'value_of_endowment_assets_at_the_beginning_of_the_fiscal_year1'
colnames(df)[dupe_vals[2]] <- 'value_of_endowment_assets_at_the_beginning_of_the_fiscal_year2'
```

Clean Data

- Combine financial columns
- Write Cleaned CSV file

Combine financial columns

```
df$endowment_total <- ifelse(is.na(df$value_of_endowment_assets_at_the_beginning_of_the_fiscal_year1),
                             df$value_of_endowment_assets_at_the_beginning_of_the_fiscal_year2,
                             df$value_of_endowment_assets_at_the_beginning_of_the_fiscal_year1)
df$finances_spent_research <- ifelse(is.na(df$research_current_year_total),
                                     df$research_total_amount,
                                     df$research_current_year_total)
df$finances_spent_student_services <- ifelse(is.na(df$student_service_total_amount),
                                             df$student_services_current_year_total,
                                             df$student_service_total_amount)
df$finances_spent_public_service <- ifelse(is.na(df$public_service_current_year_total),
                                           df$public_service_total_amount,
                                           df$public_service_current_year_total)
df$finances_spent_academic_support <- ifelse(is.na(df$academic_support_current_year_total),
                                             df$academic_support_total_amount,
                                             df$academic_support_current_year_total)
df$finances_spent_instruction <- ifelse(is.na(df$instruction_current_year_total),
                                        df$instruction_total_amount,
                                        df$instruction_current_year_total)
df$revenue_total <- ifelse(is.na(df$total_revenues_and_investment_return_total),
                           df$total_all_revenues_and_other_additions,
                           df$total_revenues_and_investment_return_total)

drop_cols <- c('value_of_endowment_assets_at_the_beginning_of_the_fiscal_year1', 'value_of_endowment_ass

df <- df[, !(colnames(df) %in% drop_cols)]
```

Remove rows with missing values on enrollment and graduation rates

```
# remove columns that have NA's
pre_df <- nrow(df)
print(paste("Total # of schools in dataset:", nrow(df)))
```

```
## [1] "Total # of schools in dataset: 2045"
```

```
remove_any_nas <- c("undergraduate_enrollment",
                    "gradrate_ba_6yrs_total"
                    )

for (i in remove_any_nas){
  df <- df %>% subset(!(is.na(df[[i]]) | (df[[i]] == 0)))
}
```

```r
# for each subgroup, if they have students from that group, they should also have a graduation rate
enroll_grad <- c("percent_of_undergraduate_enrollment_that_are_black_or_african_american" = "gradrate_ba
                 "percent_of_undergraduate_enrollment_that_are_hispanic_latino" = "gradrate_ba_6yrs_l
                 "percent_of_undergraduate_enrollment_that_are_american_indian_or_alaska_native" = "g
                 "percent_of_undergraduate_students_awarded_pell_grants" = "pell_grant_recipients_ov
               )

for (i in names(enroll_grad)){
  df <- df %>% subset(ifelse(!is.na(df[[i]]) & (df[[i]] != 0),
                             !is.na(df[[enroll_grad[[i]]]]),
                             TRUE))
}


# remove any rows if they do not have *any* students from underrepresented groups
df <- df %>% mutate(across(where(is.numeric), ~replace_na(.x, 0))) %>% mutate(sum_subgroups = percent_o

# remove any schools where there are no underrepresented subgroups
df <- df %>% subset(sum_subgroups != 0)

# compute mean
df <- df %>% mutate(mean_subgroups = (percent_of_undergraduate_enrollment_that_are_black_or_african_amer

# convert mean subgroups to quantiles
df <- df %>% mutate(diversity_quantiles = quantcut(df$mean_subgroups, q = 4, na.rm = TRUE))
df$diversity_quantiles <- mapvalues(df$diversity_quantiles,
                                    from = sort(unique(quantcut(df$mean_subgroups, q = 4, na.rm = TRUE))
                                    to = paste("diversity quantile", 1:4))

# convert the sector into the actual names
for (i in unique(value_mapping$VariableName)){
  temp_map_val <- value_mapping %>% subset(VariableName == i)
  df[[i]] <- mapvalues(df[[i]], from = temp_map_val$Value, to = temp_map_val$ValueLabel)
}
```

```
## The following `from` values were not present in `x`: 1, 2, 3

## The following `from` values were not present in `x`: 1, 2

## The following `from` values were not present in `x`: 11, 12, 13, 21, 22, 23, 31, 32, 33, 41, 42, 43

## The following `from` values were not present in `x`: 1, 2, 3, 5

## The following `from` values were not present in `x`: 1, 2, -1

## The following `from` values were not present in `x`: 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
```

```r
paste("Removed", pre_df-nrow(df), "rows with missing undergraduate enrollment and graduation rates data
```

```
## [1] "Removed 515 rows with missing undergraduate enrollment and graduation rates data."
```

```
paste("Total # of schools for analysis:", nrow(df))
```

```
## [1] "Total # of schools for analysis: 1530"
```

```
write.csv(df, "../data/all_data_merged_df(NEW).csv", row.names = FALSE)
```

Perform EDA

```
gtown <- df %>% subset(institution_name == 'Georgetown University')
gtown
```

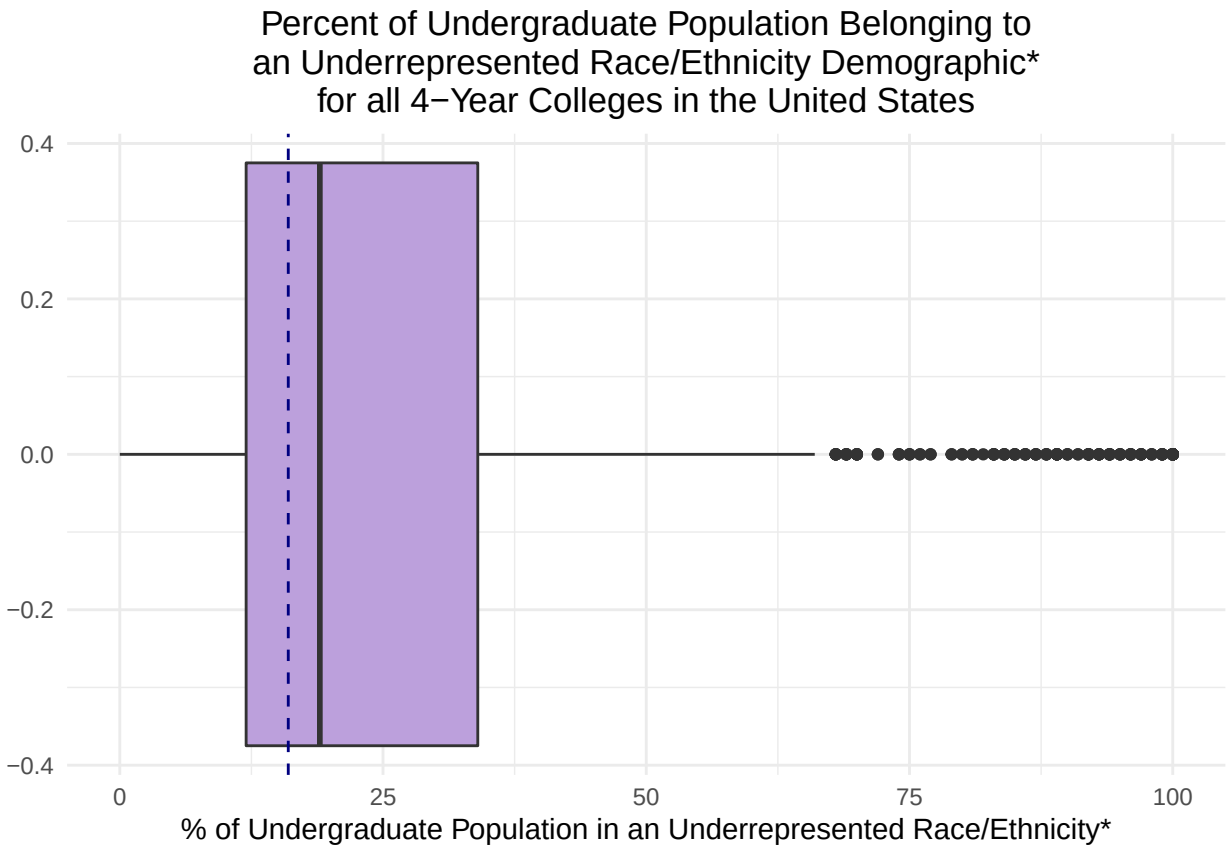
Grab Georgetown Values

```
##     unitid      institution_name
## 604 131496 Georgetown University
##     percent_of_undergraduate_students_awarded_federal_state_local_institutional_or_other_sources_of_g
## 604
##     average_amount_of_federal_state_local_institutional_or_other_sources_of_grant_aid_awarded_to_unde
## 604
##     percent_of_undergraduate_students_awarded_pell_grants
## 604                                                    13
##     average_amount_pell_grant_aid_awarded_to_undergraduate_students
## 604                                                            4670
##     percent_of_undergraduate_students_awarded_federal_student_loans
## 604                                                              25
##     average_amount_of_federal_student_loans_awarded_to_undergraduate_students
## 604                                                                      4745
##     average_net_price_students_awarded_grant_or_scholarship_aid
## 604                                                           0
##     average_net_price_income_0_30_000_students_awarded_title_iv_federal_financial_aid
## 604                                                                               0
##     average_net_price_income_over_110_000_students_awarded_title_iv_federal_financial_aid
## 604                                                                                   0
##     published_in_state_tuition_and_fees published_out_of_state_tuition_and_fees
## 604                               54104                                  54104
##     off_campus_not_with_family_room_and_board on_campus_room_and_board
## 604                                         0                   15850
##     total_price_for_in_state_students_living_on_campus
## 604                                              73840
##     total_price_for_out_of_state_students_living_on_campus
## 604                                                  73840
##     total_price_for_in_state_students_living_off_campus_not_with_family
## 604                                                                   0
##     total_price_for_out_of_state_students_living_off_campus_not_with_family
## 604                                                                       0
##     gradrate_ba_6yrs_total gradrate_ba_6yrs_men gradrate_ba_6yrs_women
## 604                     94                   93                     95
##     gradrate_ba_6yrs_black_non_hispanic gradrate_ba_6yrs_hispanic
## 604                                  93                       91
```

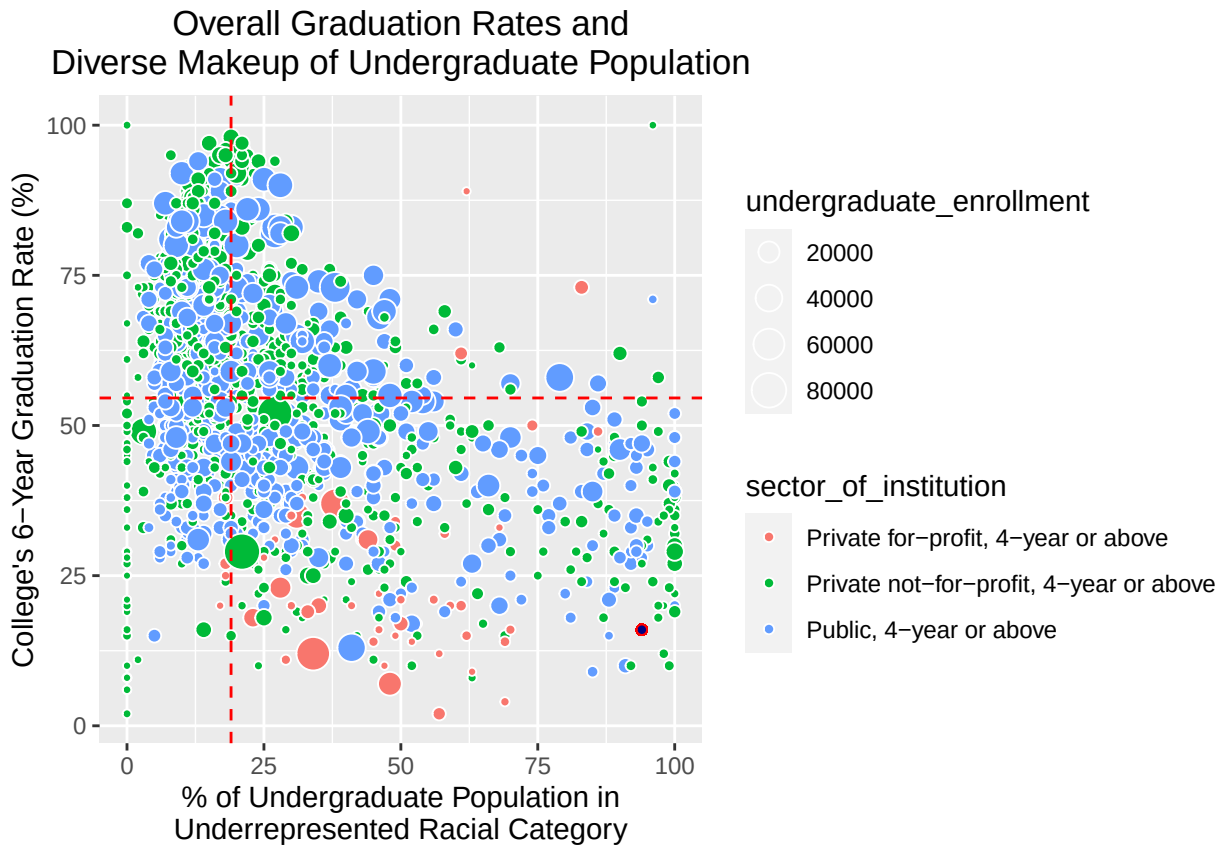
```
##     gradrate_ba_6yrs_white_non_hispanic
## 604                                 95
##     pell_grant_recipients_overall_graduation_rate_within_150_percent_of_normal_time
## 604                                                                             94
##     subsidized_stafford_loan_recipients_not_receiving_pell_grants_overall_graduation_rate_within_150_
## 604
##     did_not_receive_pell_grants_or_subsidized_stafford_loans_overall_graduation_rate_within_150_perc
## 604
##     historically_black_college_or_university
## 604                                       No
##                       sector_of_institution
## 604 Private not-for-profit, 4-year or above
##                              institutional_category
## 604 Degree-granting, primarily baccalaureate or above
##     degree_of_urbanization_urban_centric_locale
## 604                                City: Large
##     carnegie_classification_2018_size_and_setting full_time_retention_rate
## 604         Four-year, large, highly residential                        96
##     part_time_retention_rate undergraduate_enrollment
## 604                      100                     7459
##     percent_of_undergraduate_enrollment_that_are_black_or_african_american
## 604                                                                     6
##     percent_of_undergraduate_enrollment_that_are_hispanic_latino
## 604                                                           10
##     percent_of_undergraduate_enrollment_that_are_white
## 604                                                50
##     percent_of_undergraduate_enrollment_that_are_women
## 604                                                56
##     endowment_assets_year_end_per_fte_enrollment
## 604                                            0
##     number_of_branches_and_independent_libraries
## 604                                            6
##     all_programs_offered_completely_via_distance_education
## 604                                                     No
##     percent_of_undergraduate_students_enrolled_exclusively_in_distance_education_courses
## 604                                                                                   0
##     percent_of_undergraduate_enrollment_that_are_asian
## 604                                                 9
##     percent_of_undergraduate_enrollment_that_are_american_indian_or_alaska_native
## 604                                                                             0
##     percent_of_undergraduate_enrollment_that_are_native_hawaiian_or_other_pacific_islander
## 604                                                                                     0
##     gradrate_ba_6yrs_american_indian_or_alaska_native gradrate_ba_6yrs_asian
## 604                                               100                    95
##     gradrate_ba_6yrs_native_hawaiian_or_other_pacific_islander endowment_total
## 604                                                          0      1769557000
##     finances_spent_research finances_spent_student_services
## 604               235133000                       137261000
##     finances_spent_public_service finances_spent_academic_support
## 604                      13593000                       193224000
##     finances_spent_instruction revenue_total sum_subgroups sum_race_subgroups
## 604                  525441000    1395067000            29                 16
##     mean_subgroups  diversity_quantiles
## 604           7.25 diversity quantile 1
```

Underrepresented Students Distribution

```
# mean of underrepresneted group makeup of the populations size histogram
# df %>% ggplot(aes(x=mean_subgroups)) + geom_boxplot(fill = "#bca0dc") + ggtitle("Percent of Undergrad

# sum of all underrepresented races = black, hispanic, native american, and pacific islander
df %>% ggplot(aes(x=sum_race_subgroups)) + geom_boxplot(fill = "#bca0dc") + ggtitle("Percent of Undergra
```




```
# does racial diversity lead to overall higher graduation rates?
ggplot(df, aes(y=gradrate_ba_6yrs_total, x=sum_race_subgroups, color = sector_of_institution)) +
  geom_point(aes(size=undergraduate_enrollment, fill = sector_of_institution), colour="white",pch=21) +
```



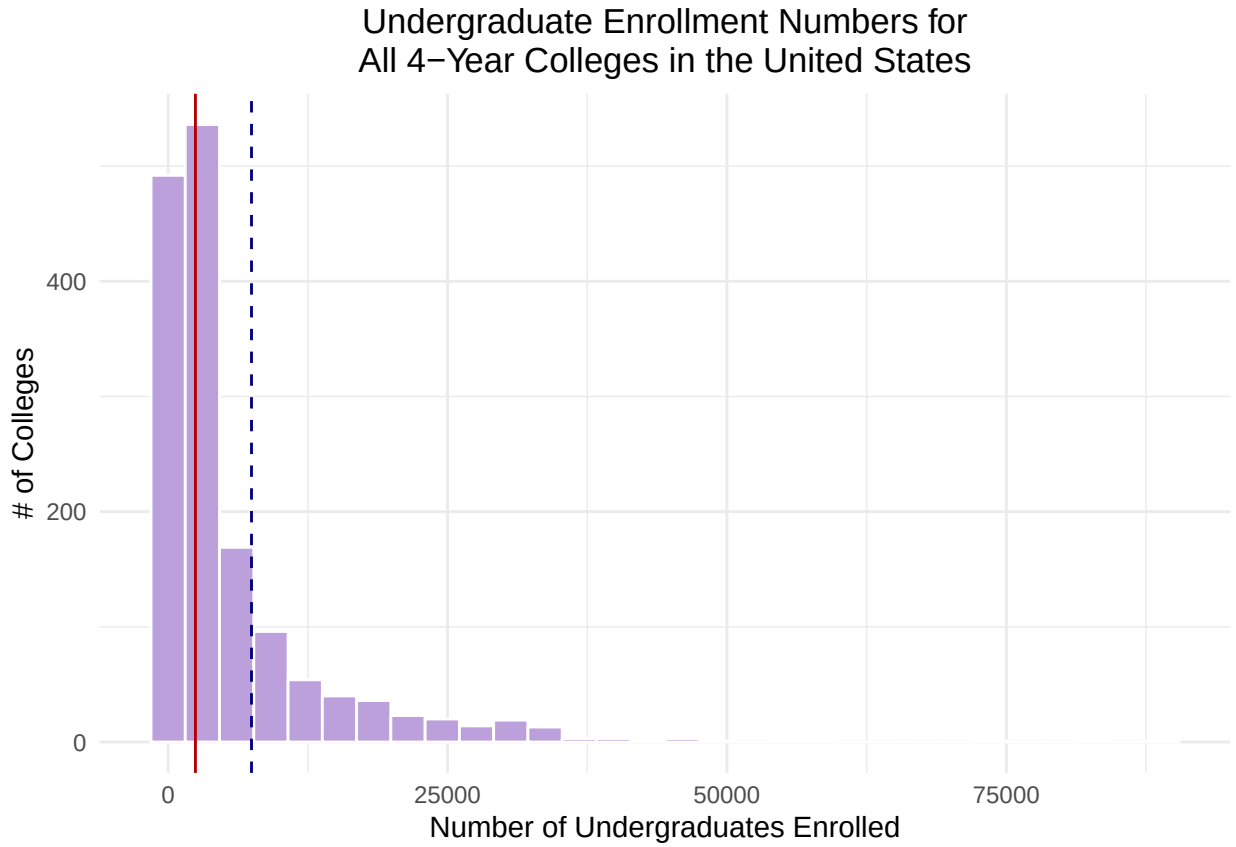

```
# geom_boxplot(aes(x=variable, y=value, fill=variable))

head(df %>% arrange(desc(sum_subgroups)) %>% select(institution_name, sum_subgroups))
```

```
##                             institution_name sum_subgroups
## 1 Universidad Ana G. Mendez-Carolina Campus           200
## 2    Universidad Ana G. Mendez-Cupey Campus           200
## 3   Universidad Ana G. Mendez-Gurabo Campus           200
## 4               Atlantic University College           196
## 5                 Dewey University-Hato Rey           196
## 6                      CEM College-Humacao            194
```

Size of College (by Undergraduate Enrollment)

```
# enrollemnt size histogram
df %>% ggplot(aes(x=undergraduate_enrollment)) + geom_histogram(position="stack", col = "white", fill =
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Look into Outliers

- After googling each of the largest schools, we see that their abnormal size is due to being primarily online institutions. We decided to keep these institutions because it would be informative to know if primarily online institutions graduate underrepresented students at higher rates. Decided to take the logarithm of the enrollment size in order to coerce the distribution into being more normal

```
# look into outliers
df %>% subset(undergraduate_enrollment > 40000) %>% select(institution_name, undergraduate_enrollment, a
```

```
##                            institution_name undergraduate_enrollment
## 1                Western Governors University                     88921
## 2           Southern New Hampshire University                     82693
## 3               University of Phoenix-Arizona                     74061
## 4               University of Central Florida                     58821
## 5                     Grand Canyon University                     54139
## 6      Texas A & M University-College Station                     53743
## 7            Florida International University                     48818
## 8        University of Maryland Global Campus                     47253
## 9           Ohio State University-Main Campus                     46820
## 10                         Liberty University                     45935
## 11 Arizona State University Campus Immersion                     42844
## 12             Brigham Young University-Idaho                     42341
## 13          The University of Texas at Austin                     40804
##    all_programs_offered_completely_via_distance_education
```

```
## 1                                                    Yes
## 2                                                     No
## 3                                                     No
## 4                                                     No
## 5                                                     No
## 6                                                     No
## 7                                                     No
## 8                                                     No
## 9                                                     No
## 10                                                    No
## 11                                                    No
## 12                                                    No
## 13                                                    No
##    percent_of_undergraduate_students_enrolled_exclusively_in_distance_education_courses
## 1                                                                                   100
## 2                                                                                    93
## 3                                                                                    99
## 4                                                                                    12
## 5                                                                                    67
## 6                                                                                     0
## 7                                                                                    19
## 8                                                                                    77
## 9                                                                                     2
## 10                                                                                   71
## 11                                                                                    0
## 12                                                                                   50
## 13                                                                                    0
```

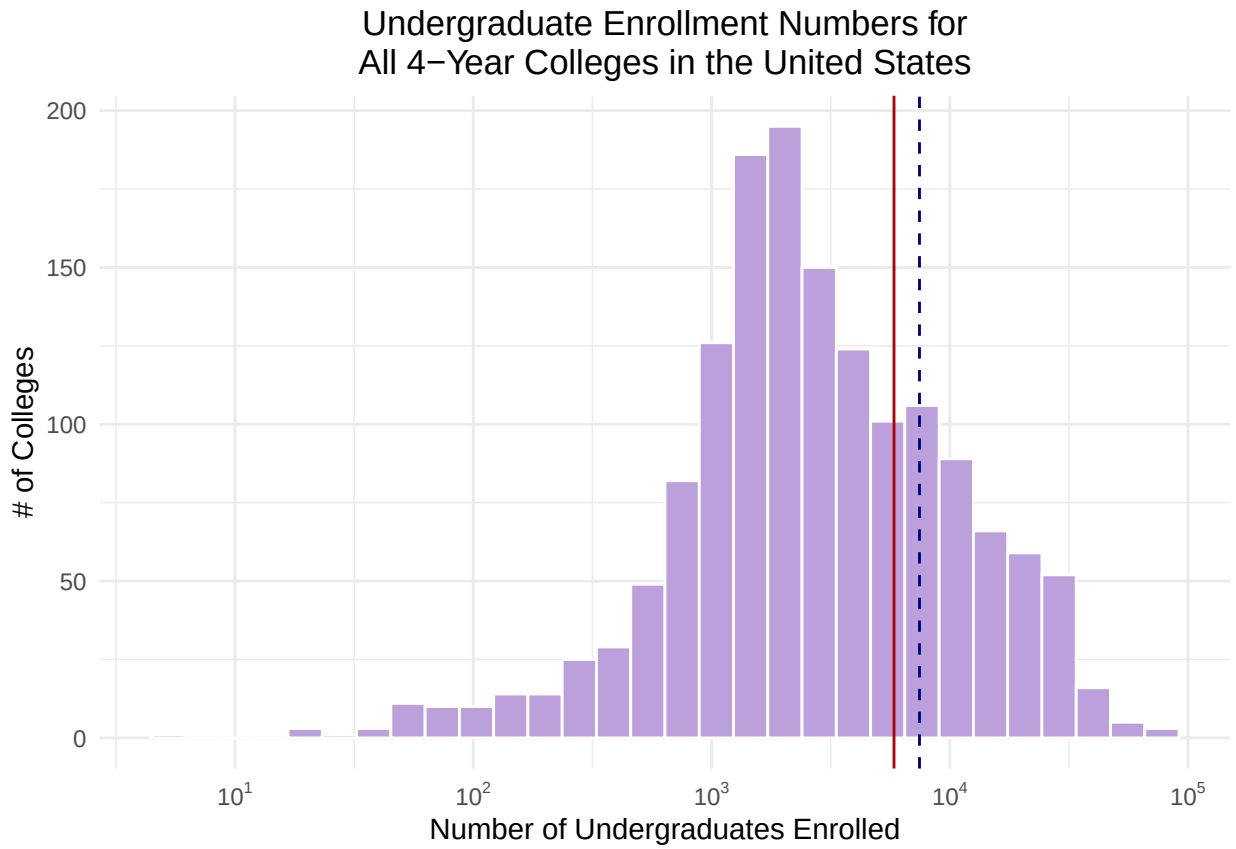
```
# found out that these are all online colleges. question - would we want to include entirely online col
```

```
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

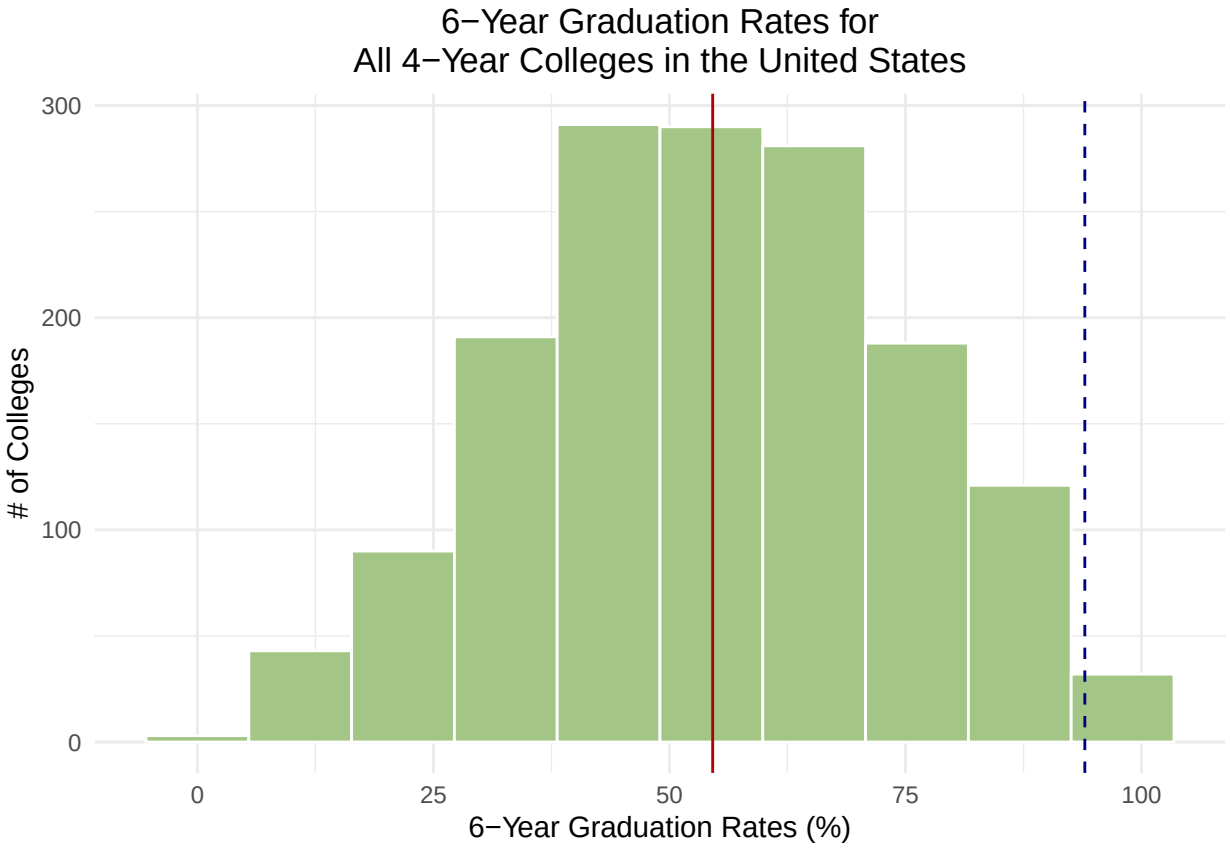
```
# redo histogram by taking log enrollment
df %>% ggplot(aes(x=undergraduate_enrollment)) + scale_x_log10(breaks = trans_breaks("log10", function(
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



6-Year Graduation Rates

```
# graduation rates histogram
df %>% ggplot(aes(x=gradrate_ba_6yrs_total)) + geom_histogram(position="stack", col = "white", fill = "
```



Look into Outliers

```
# look out graduate rate outliers

df %>% subset(gradrate_ba_6yrs_total < 10) %>% select(institution_name, undergraduate_enrollment, secto
```

```
##                            institution_name undergraduate_enrollment
## 86                          Ashford University                    28701
## 110                            Bacone College                      271
## 158  Beth Hamedrash Shaarei Yosher Institute                       74
## 653             Harris-Stowe State University                     1716
## 1201            Rabbinical College of America                      224
## 1316          Sh'or Yoshuv Rabbinical College                      138
## 1346              South University-Montgomery                      321
## 1349         South University-Savannah Online                     3568
## 1476            Talmudical Academy-New Jersey                       58
## 1483                   Telshe Yeshiva-Chicago                       77
## 1773           University of Phoenix-Illinois                       51
## 2024             Yeshiva of the Telshe Alumni                      101
## 2038                       Yeshivas Novominsk                      159
##                        sector_of_institution
## 86       Private for-profit, 4-year or above
## 110  Private not-for-profit, 4-year or above
## 158  Private not-for-profit, 4-year or above
## 653                  Public, 4-year or above
## 1201 Private not-for-profit, 4-year or above
## 1316 Private not-for-profit, 4-year or above
```

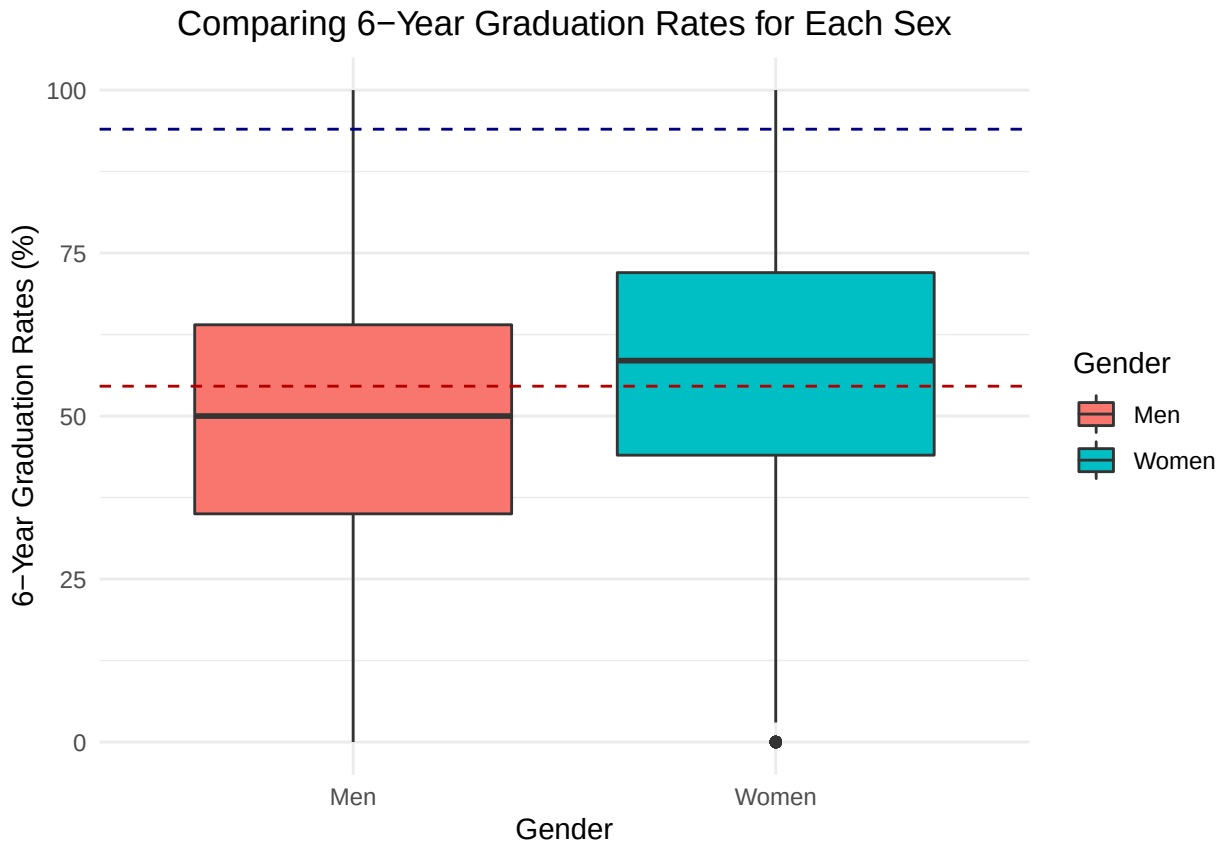
```
## 1346     Private for-profit, 4-year or above
## 1349     Private for-profit, 4-year or above
## 1476 Private not-for-profit, 4-year or above
## 1483 Private not-for-profit, 4-year or above
## 1773     Private for-profit, 4-year or above
## 2024 Private not-for-profit, 4-year or above
## 2038 Private not-for-profit, 4-year or above
##                                   institutional_category
## 86       Degree-granting, primarily baccalaureate or above
## 110      Degree-granting, primarily baccalaureate or above
## 158      Degree-granting, primarily baccalaureate or above
## 653      Degree-granting, primarily baccalaureate or above
## 1201     Degree-granting, primarily baccalaureate or above
## 1316     Degree-granting, primarily baccalaureate or above
## 1346     Degree-granting, primarily baccalaureate or above
## 1349     Degree-granting, primarily baccalaureate or above
## 1476     Degree-granting, primarily baccalaureate or above
## 1483     Degree-granting, primarily baccalaureate or above
## 1773     Degree-granting, primarily baccalaureate or above
## 2024 Degree-granting, not primarily baccalaureate or above
## 2038     Degree-granting, primarily baccalaureate or above
##      gradrate_ba_6yrs_total gradrate_ba_6yrs_black_non_hispanic
## 86                        7                                   5
## 110                       8                                   9
## 158                       6                                   0
## 653                       9                                   9
## 1201                      8                                   0
## 1316                      8                                   0
## 1346                      4                                   5
## 1349                      2                                   2
## 1476                      8                                   0
## 1483                      6                                   0
## 1773                      9                                   0
## 2024                      6                                   0
## 2038                      2                                   0
##      gradrate_ba_6yrs_white_non_hispanic
## 86                                     9
## 110                                    6
## 158                                    6
## 653                                   20
## 1201                                  11
## 1316                                   8
## 1346                                   0
## 1349                                   3
## 1476                                   8
## 1483                                   6
## 1773                                  50
## 2024                                   6
## 2038                                   2
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

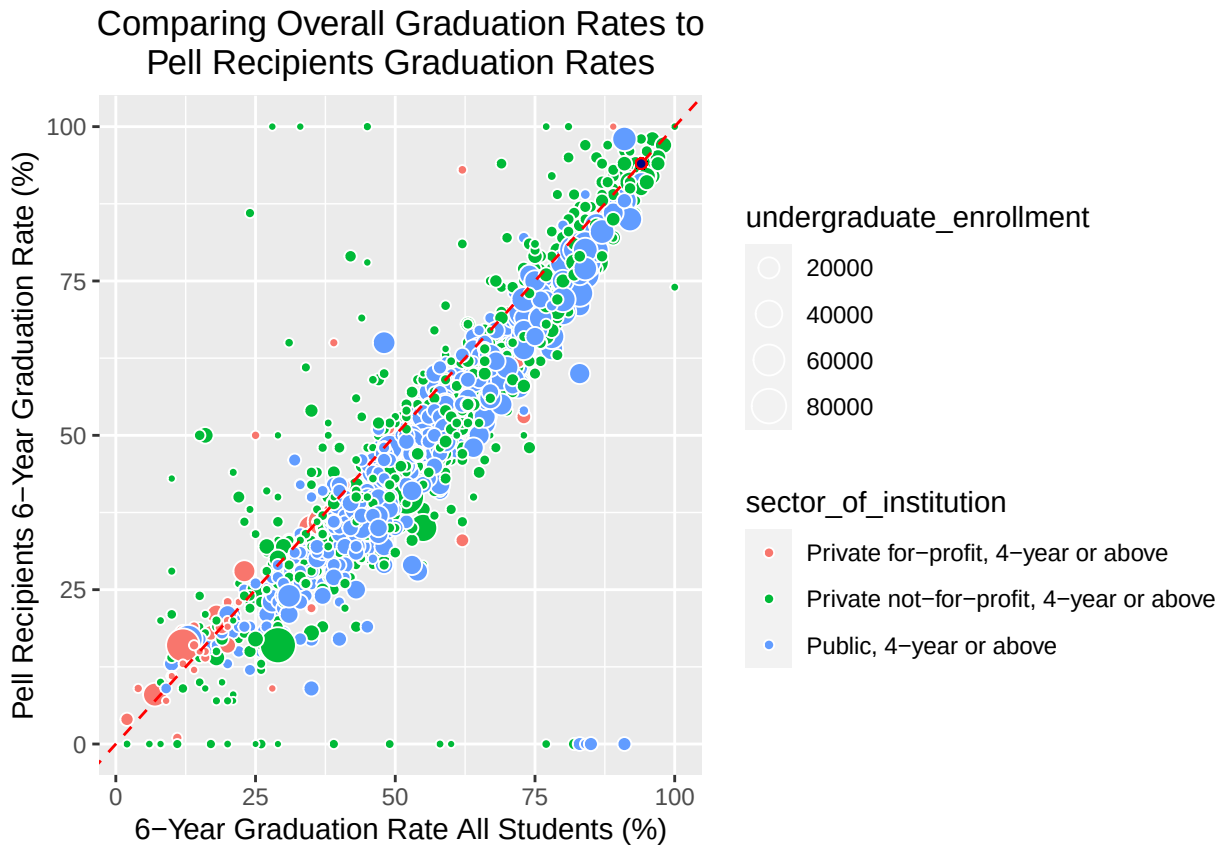
```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggplot2)

ggplot(melt(rename(df, Men = gradrate_ba_6yrs_men, Women = gradrate_ba_6yrs_women), id.vars = 'unitid',
```



Comparing Graduation Rates for Different Subgroups of Students

```
# any points above the red dotted lines are schools where Pell-Grant recipients are graduating at *high
# we added the dimension of school size to the mix and as you can see, majority of the schools where Pe
ggplot(df, aes(x=gradrate_ba_6yrs_total, y=pell_grant_recipients_overall_graduation_rate_within_150_per
  geom_point(aes(size=undergraduate_enrollment, fill = sector_of_institution), colour="white",pch=21) +
```

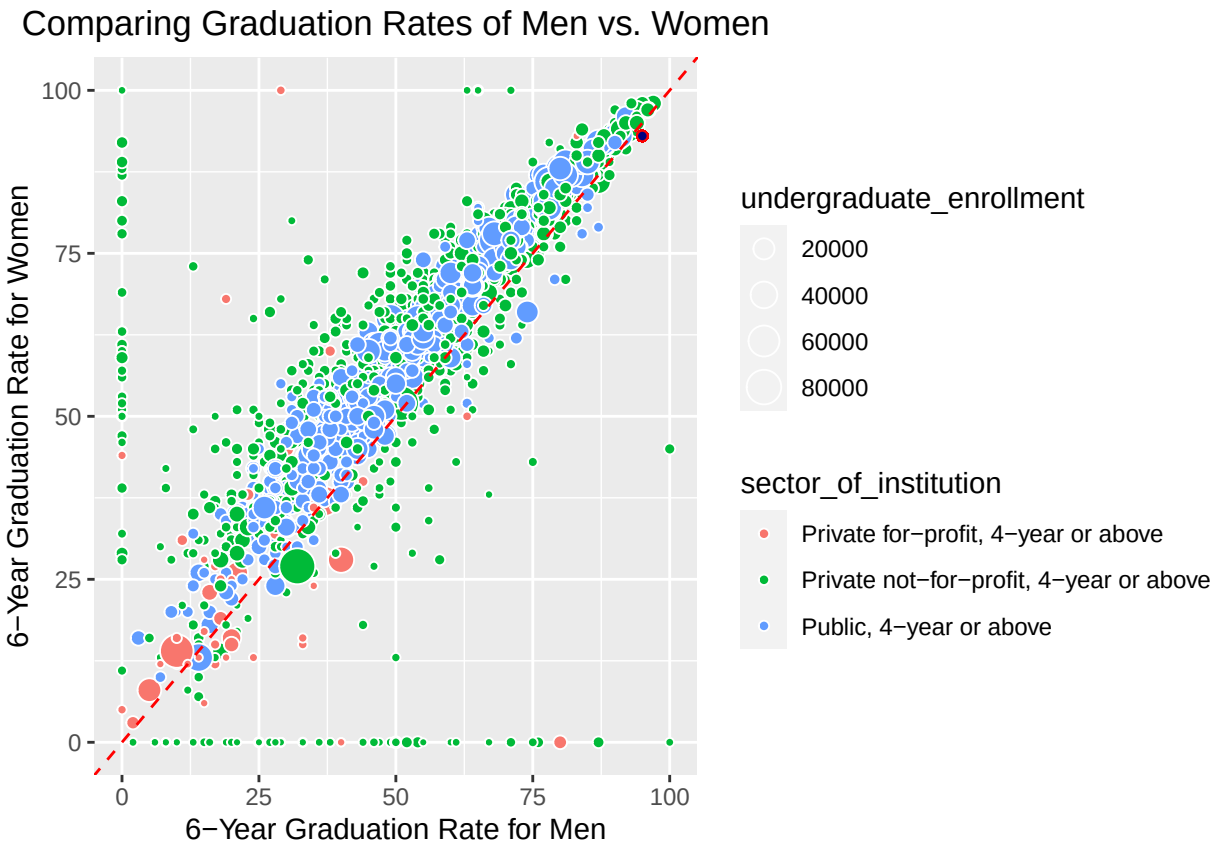



```
paste("There are", sum(df$pell_grant_recipients_overall_graduation_rate_within_150_percent_of_normal_ti
```

```
## [1] "There are 225 schools where Pell-Grant recipients graduate at higher rates than that school's o
```

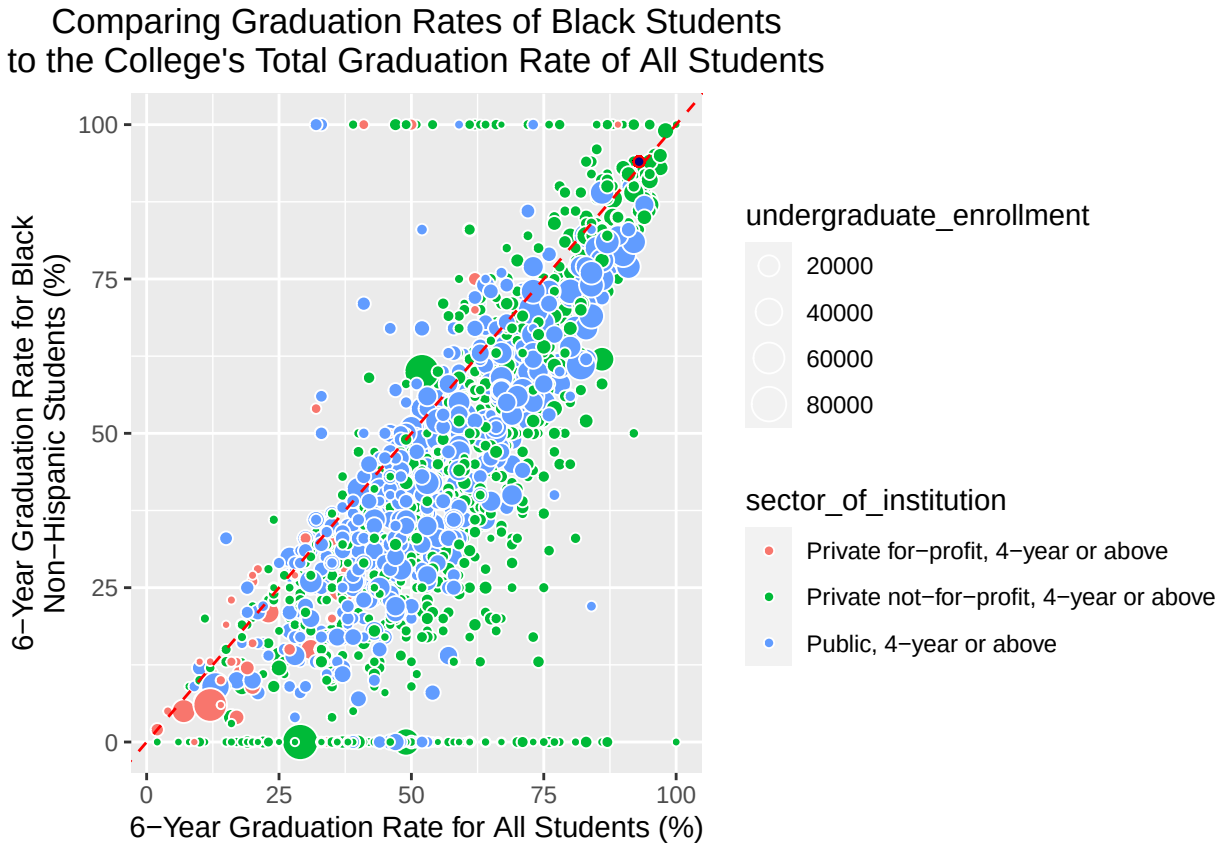
```
# create a derived field that indicates whether pell students are doing about the same or better than o
df <- df %>% mutate(pell_above_avg = pell_grant_recipients_overall_graduation_rate_within_150_percent_o
```

```
# any points above the red dotted lines are schools where women graduate at higher rates than men
ggplot(df, aes(y=gradrate_ba_6yrs_women, x=gradrate_ba_6yrs_men, color = sector_of_institution)) +
  geom_point(aes(size=undergraduate_enrollment, fill = sector_of_institution), colour="white",pch=21) +
```




```
# create a derived field that indicates whether women and men are graduating at the same rates
df <- df %>% mutate(women_above_avg = gradrate_ba_6yrs_women >= gradrate_ba_6yrs_men)
```

```
# any points above the red dotted lines are schools where black students graduate at higher rates than
ggplot(df, aes(y=gradrate_ba_6yrs_black_non_hispanic, x=gradrate_ba_6yrs_total)) +
  geom_point(aes(size=undergraduate_enrollment, fill = sector_of_institution), colour="white", pch=21) +
```

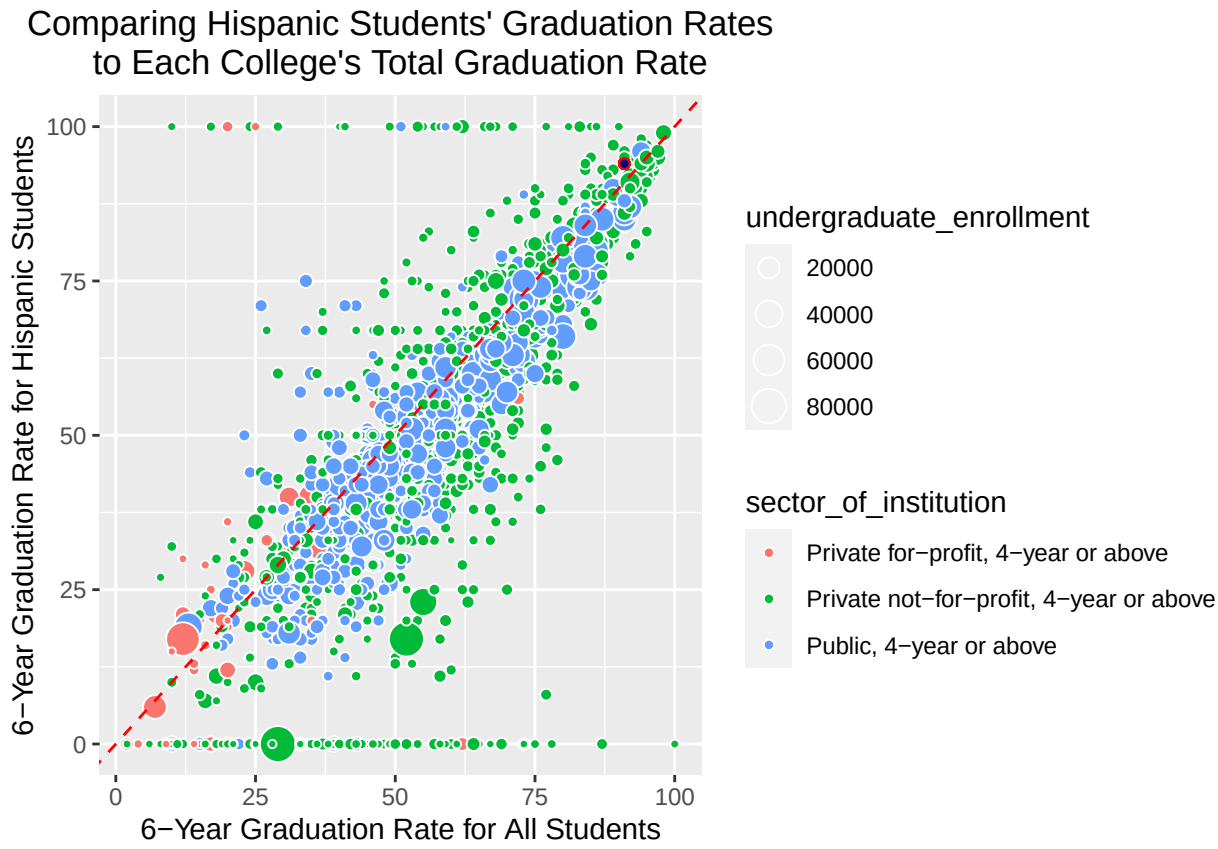



```
paste("There are", sum(df$gradrate_ba_6yrs_black_non_hispanic > df$gradrate_ba_6yrs_total), "schools wh
```

```
## [1] "There are 207 schools where Black students graduate at higher rates than White students."
```

```
# create a derived field that indicates whether black students are graduating at the same rates at the
df <- df %>% mutate(black_above_avg = gradrate_ba_6yrs_black_non_hispanic >= gradrate_ba_6yrs_total)
```

```
# any points above the red dotted lines are schools where hispanic students graduate at higher rates th
ggplot(df, aes(y=gradrate_ba_6yrs_hispanic, x=gradrate_ba_6yrs_total, color = sector_of_institution)) +
  geom_point(aes(size=undergraduate_enrollment, fill = sector_of_institution), colour="white",pch=21) +
```

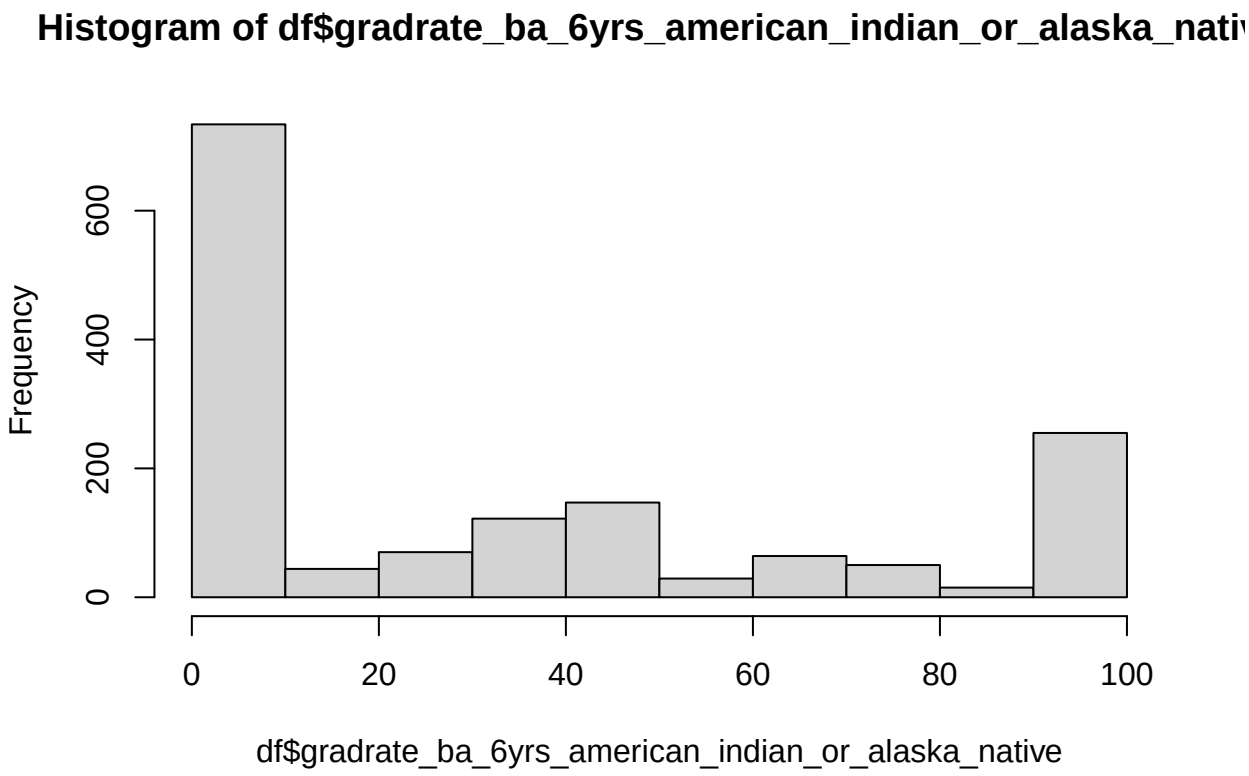



```
paste("There are", sum(df$gradrate_ba_6yrs_hispanic > df$gradrate_ba_6yrs_total), "schools where Hispani
```

```
## [1] "There are 385 schools where Hispanic/Latinx students graduate at higher rates than White studen
```

```
# create a derived field that indicates whether hispanic students are graduating at the same rates at t
df <- df %>% mutate(hispanic_above_avg = gradrate_ba_6yrs_hispanic >= gradrate_ba_6yrs_total)
```

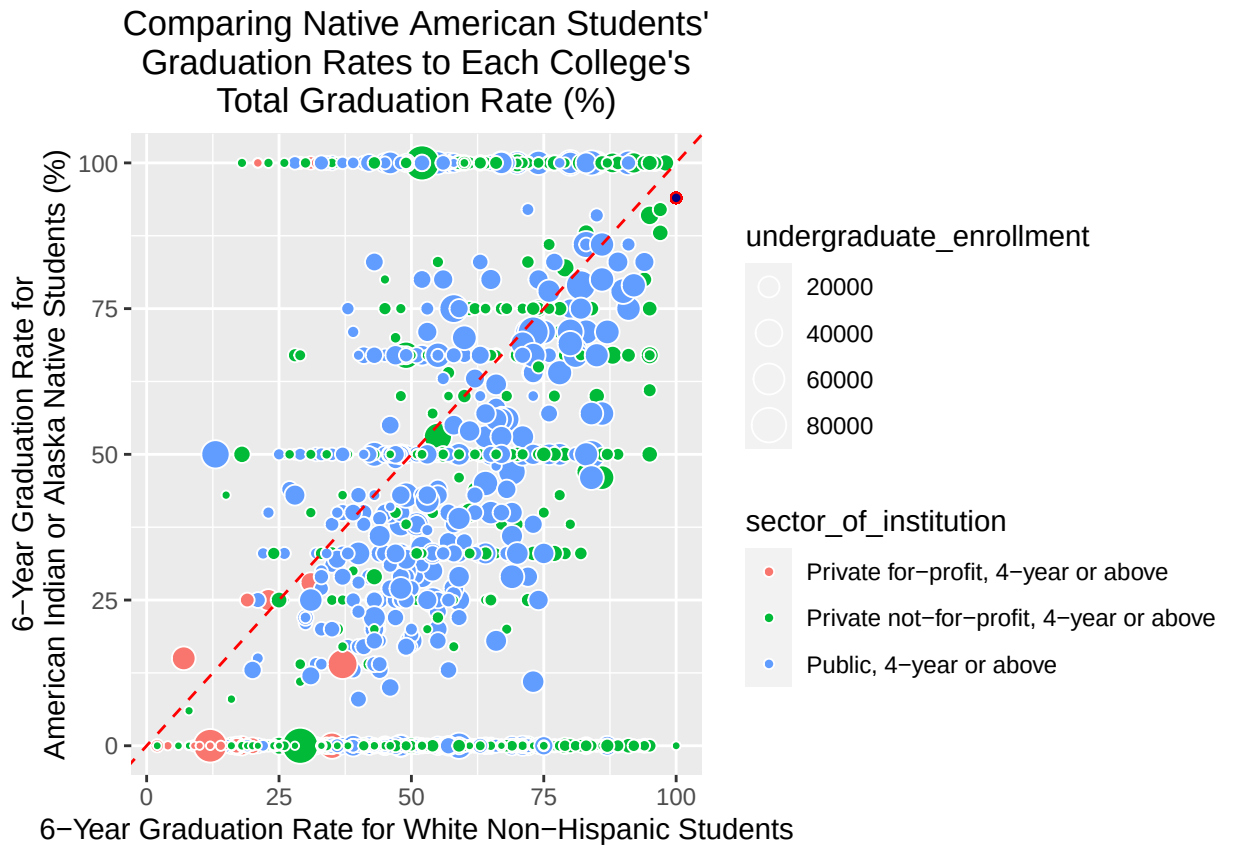
```
hist(df$gradrate_ba_6yrs_american_indian_or_alaska_native)
```




```
print(sum(!is.na(df$gradrate_ba_6yrs_american_indian_or_alaska_native)))
```

```
## [1] 1530
```

```
# any points above the red dotted lines are schools where hispanic students graduate at higher rates th
ggplot(df, aes(y=gradrate_ba_6yrs_american_indian_or_alaska_native, x=gradrate_ba_6yrs_total, color = s
```




```
paste("There are", sum(df$gradrate_ba_6yrs_american_indian_or_alaska_native > df$gradrate_ba_6yrs_total
```

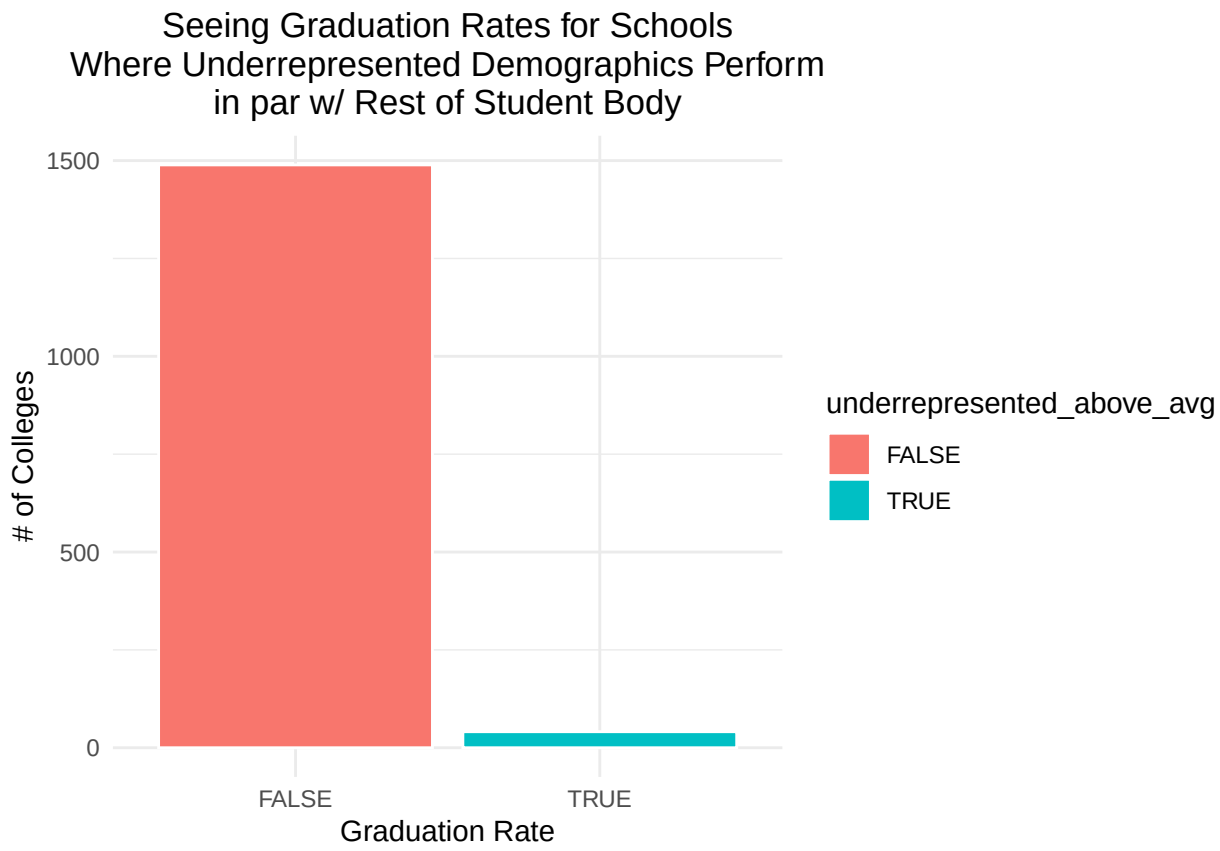
```
## [1] "There are 394 schools where Native American students graduate at higher rates than White studen
```

Dive into Schools where Underrepresented Student Success Rates Mirror or Surpass General Student Body

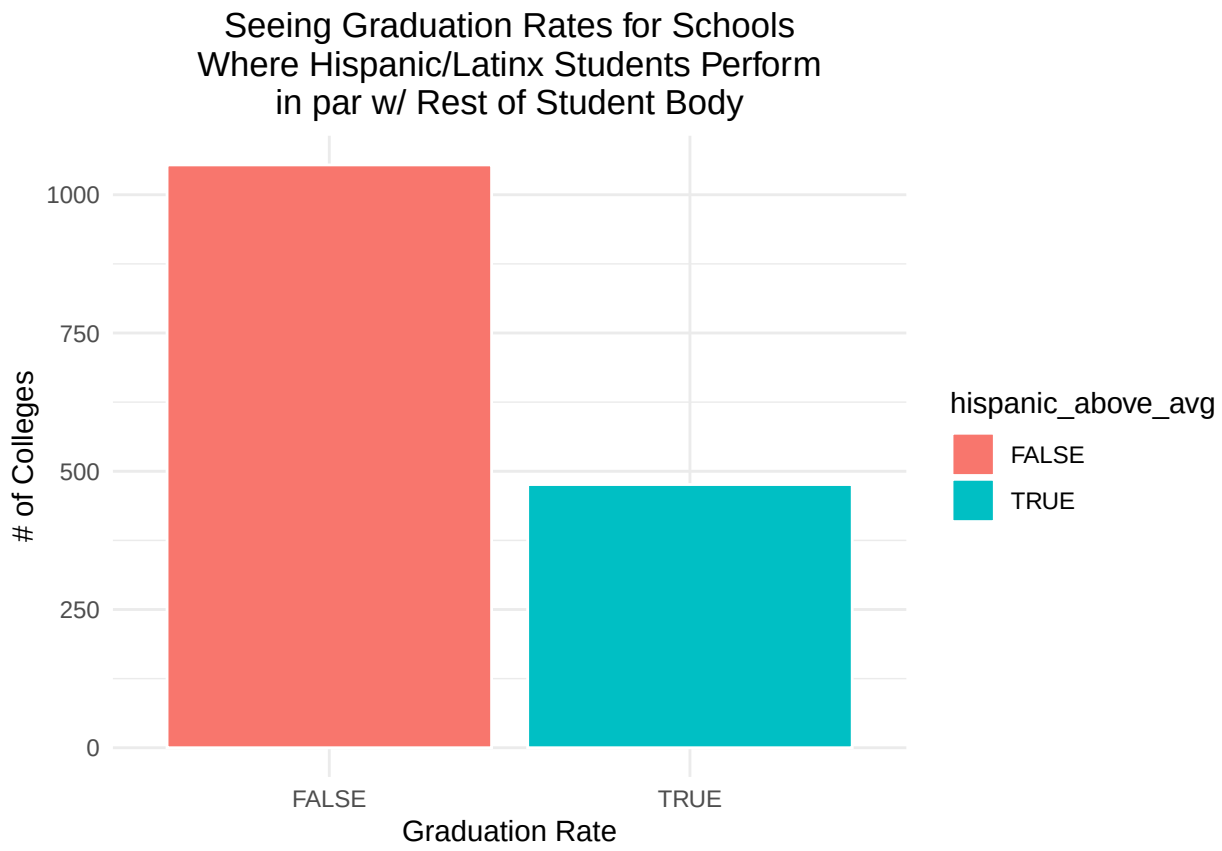
```
# see how many schools are ones in which underrepresented students are doing the same or better as the
df <- df %>% mutate(underrepresented_above_avg = hispanic_above_avg & pell_above_avg & women_above_avg &

print(table(df$underrepresented_above_avg))
```

```
##
## FALSE  TRUE
##  1489    41
```

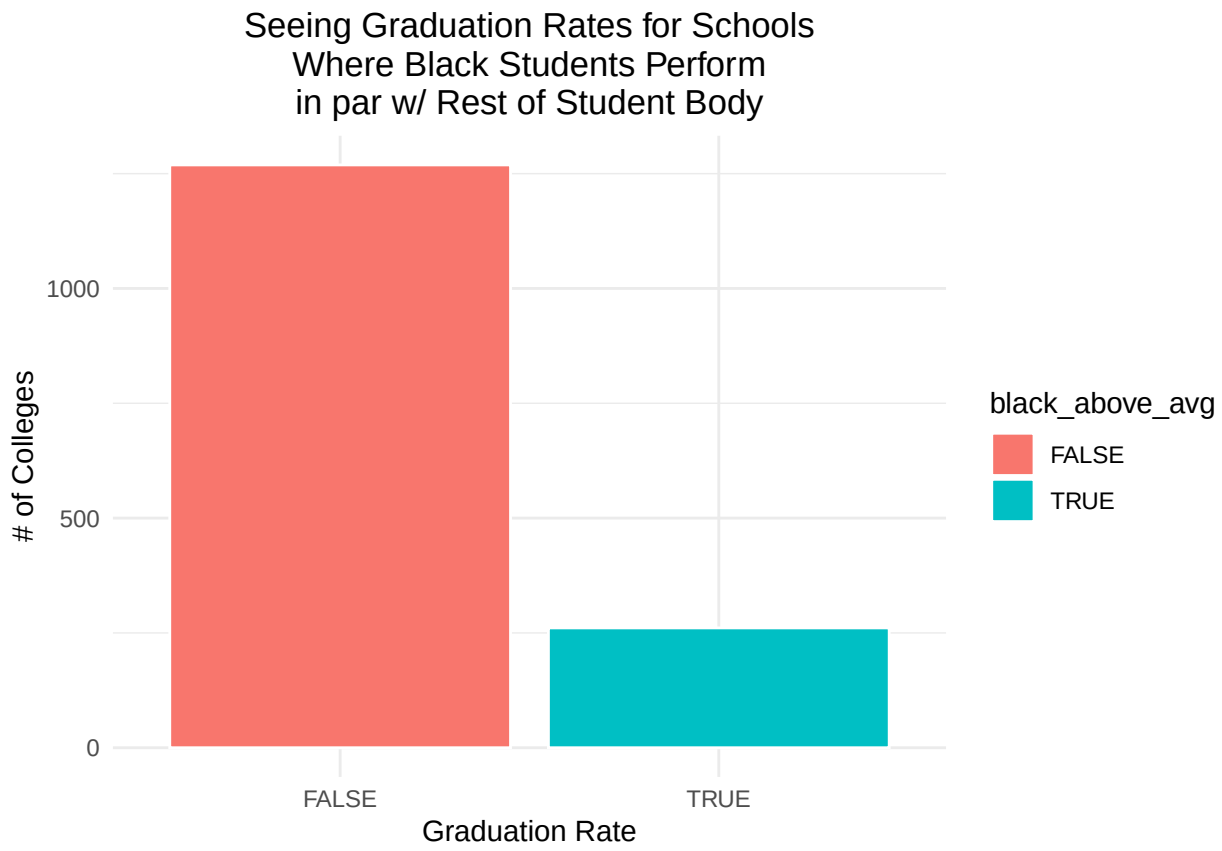
```
subgroups_success_df <- df %>% subset(underrepresented_above_avg==TRUE)

# df %>% subset(underrepresented_above_avg == TRUE) %>% ggplot(aes(x=gradrate_ba_6yrs_total,fill = unde

df %>% ggplot(aes(x=underrepresented_above_avg, fill = underrepresented_above_avg)) + geom_bar(position=
```



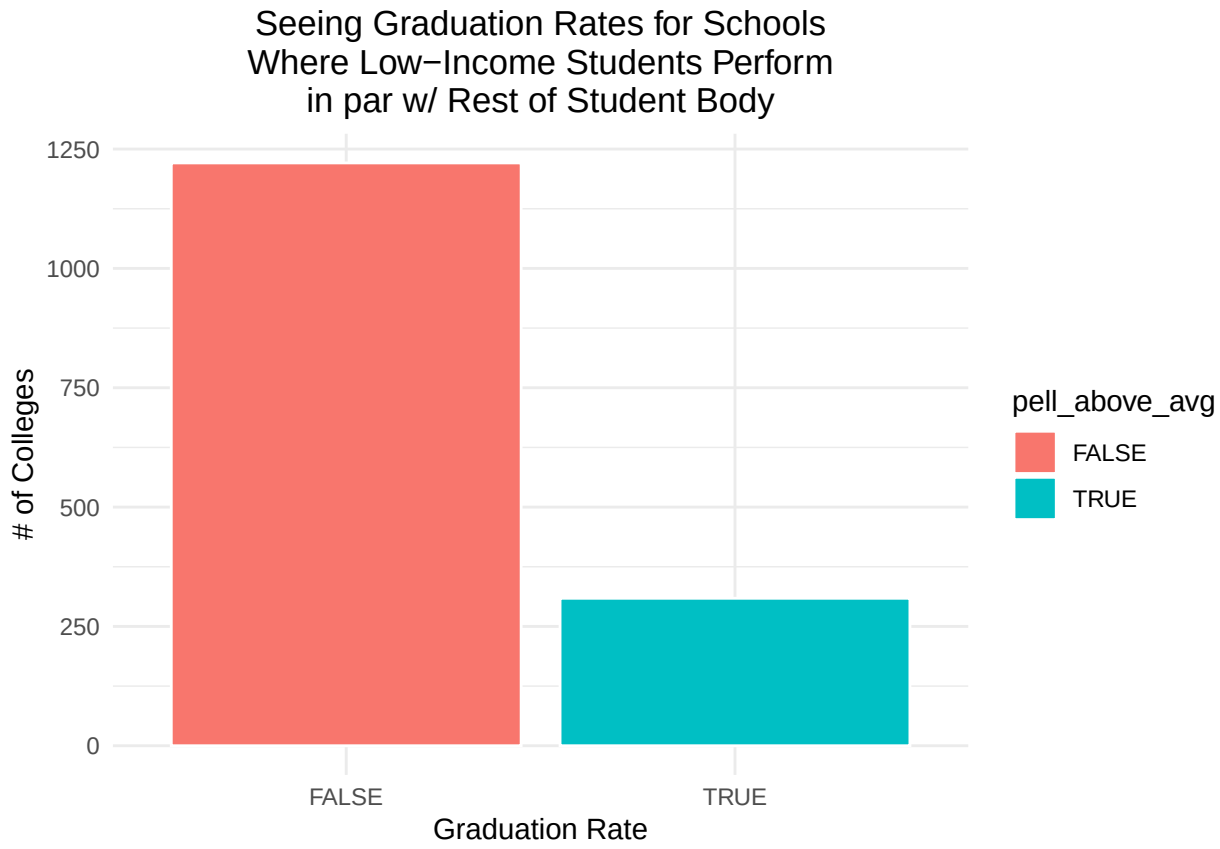

```
df %>% ggplot(aes(x=hispanic_above_avg, fill = hispanic_above_avg)) + geom_bar(position="stack", col =
```



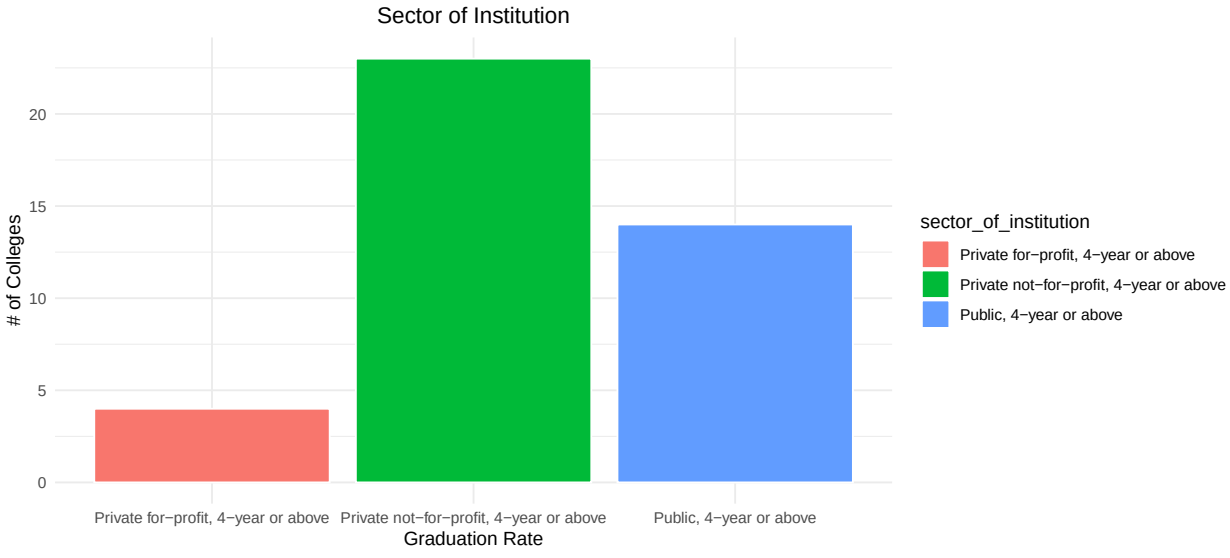

```
df %>% ggplot(aes(x=black_above_avg, fill = black_above_avg)) + geom_bar(position="stack", col = "white
```



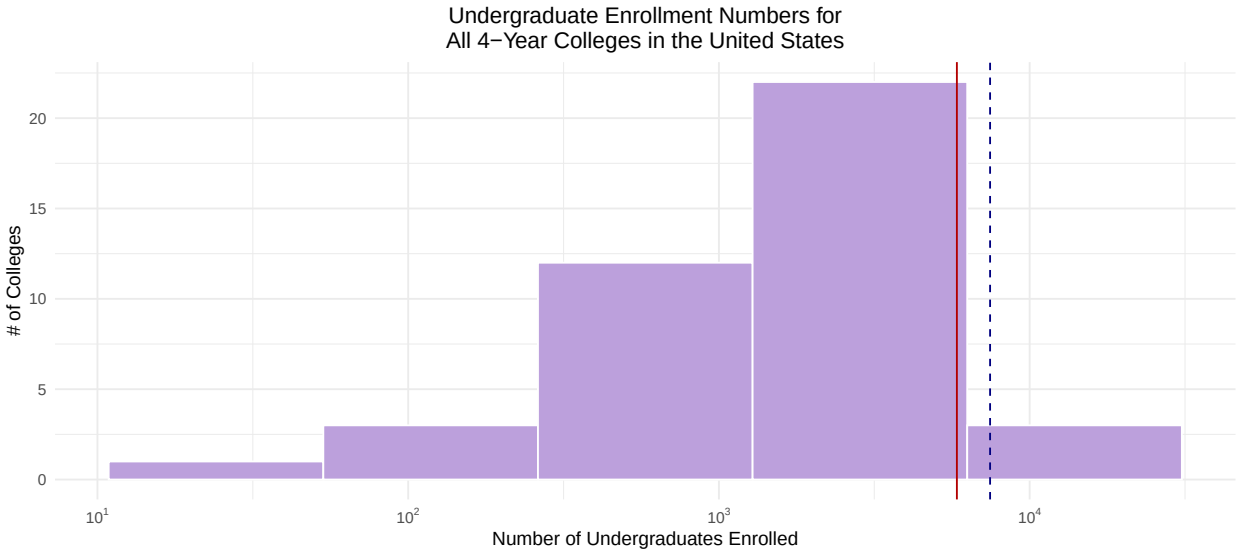

```
df %>% ggplot(aes(x=pell_above_avg, fill = pell_above_avg)) + geom_bar(position="stack", col = "white")
```




```
subgroups_success_df %>% ggplot(aes(x=sector_of_institution, fill = sector_of_institution)) + geom_bar(
```

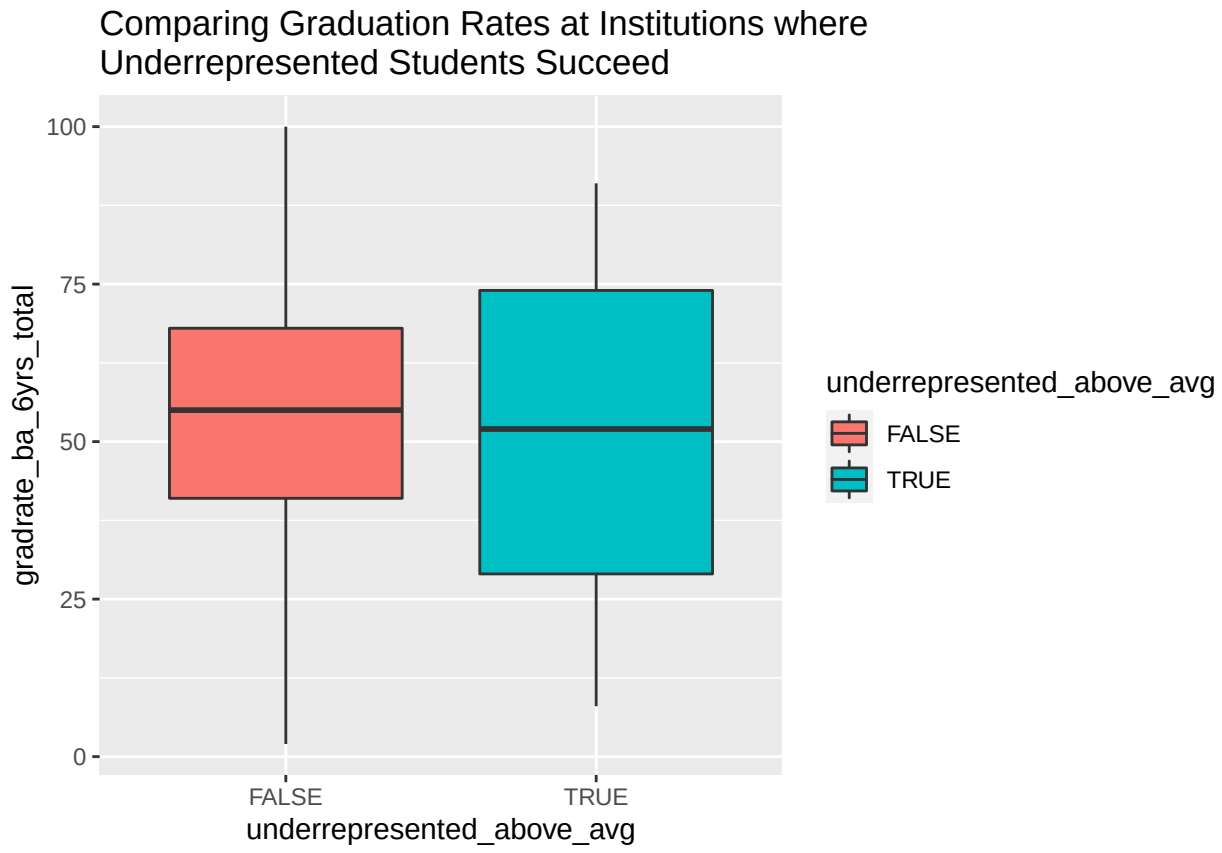


```
subgroups_success_df %>% ggplot(aes(x=undergraduate_enrollment)) + scale_x_log10(breaks = trans_breaks(
```

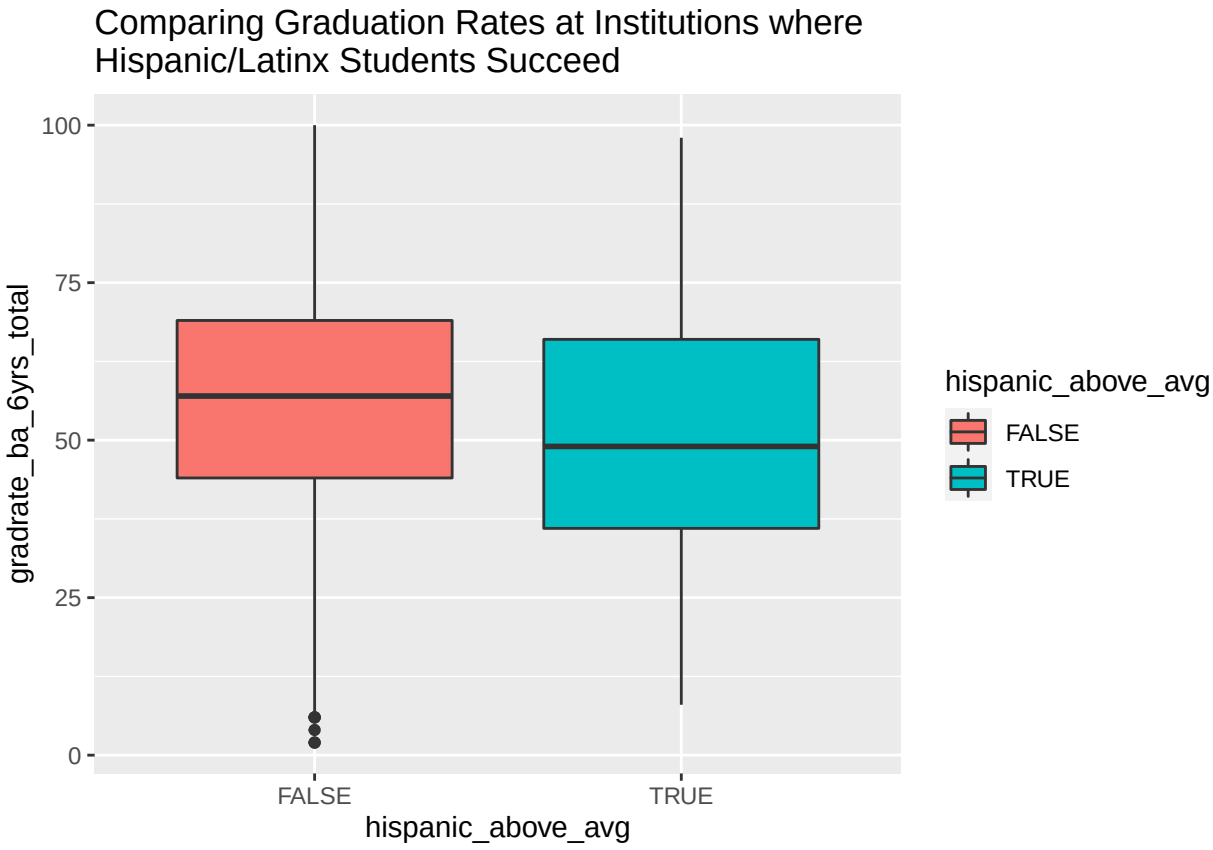


Do schools where underrepresented students do better have higher graduation rates?

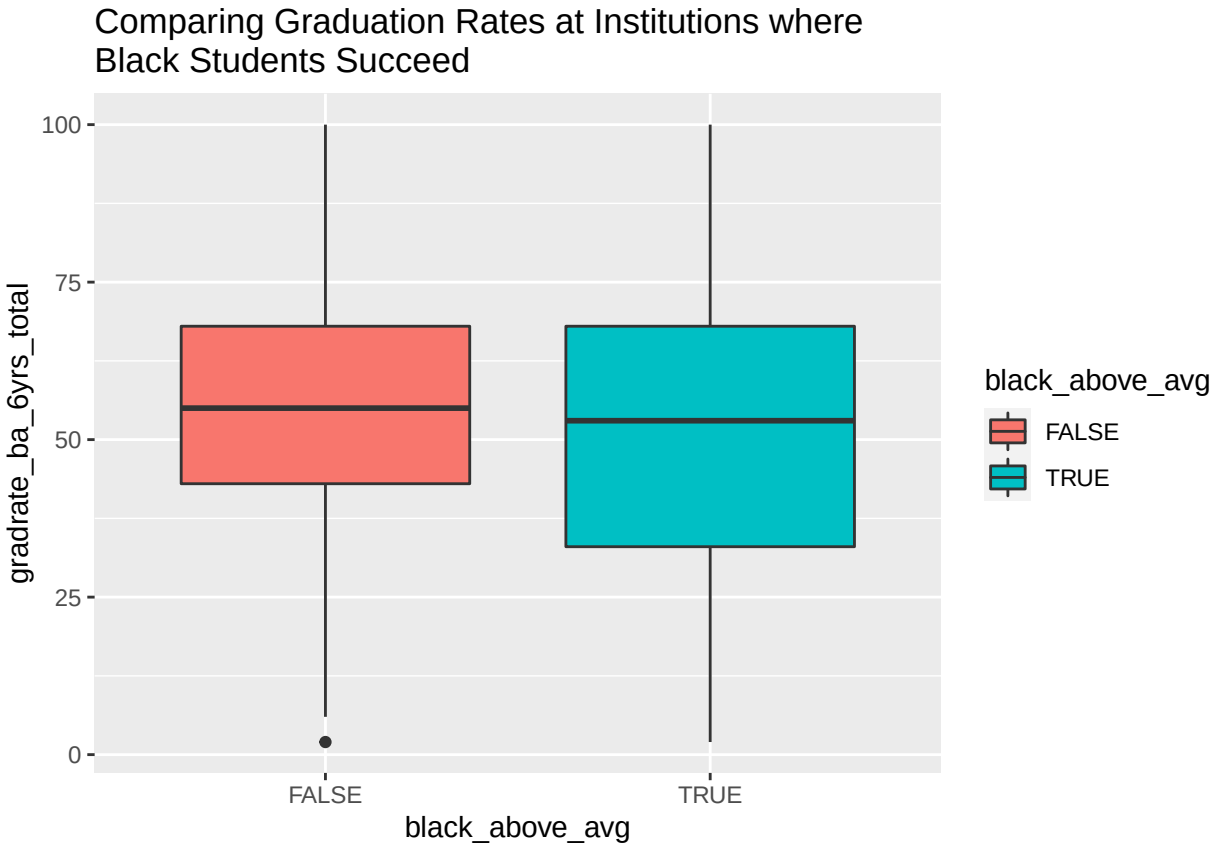
```
df %>%
  ggplot( aes(x = underrepresented_above_avg, y = gradrate_ba_6yrs_total, fill=underrepresented_above_av
  geom_boxplot() +ggtitle("Comparing Graduation Rates at Institutions where\nUnderrepresented Students S
```



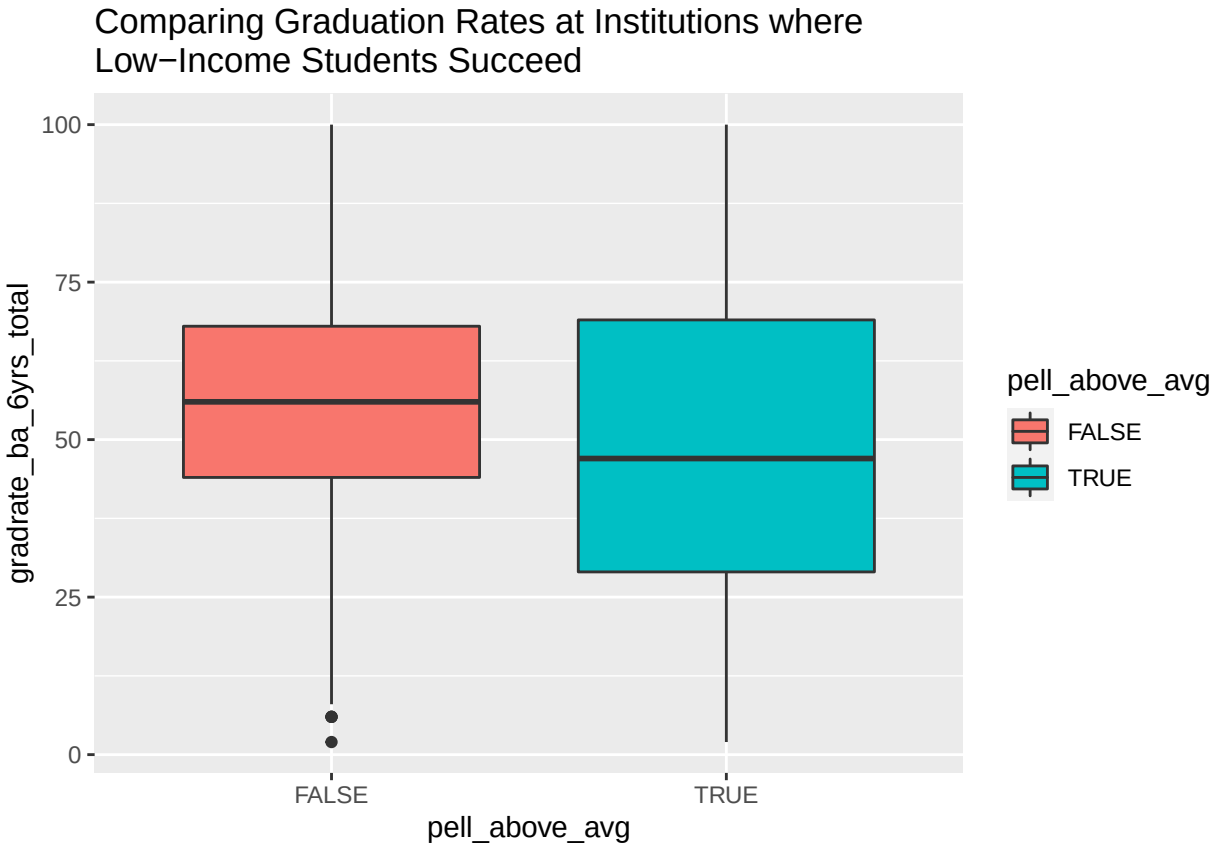
```
df %>%
  ggplot( aes(x = hispanic_above_avg, y = gradrate_ba_6yrs_total, fill=hispanic_above_avg)) +
  geom_boxplot() +ggtitle("Comparing Graduation Rates at Institutions where\nHispanic/Latinx Students Su
```




```
df %>%
  ggplot( aes(x = black_above_avg, y = gradrate_ba_6yrs_total, fill=black_above_avg)) +
  geom_boxplot() +ggtitle("Comparing Graduation Rates at Institutions where\nBlack Students Succeed")
```

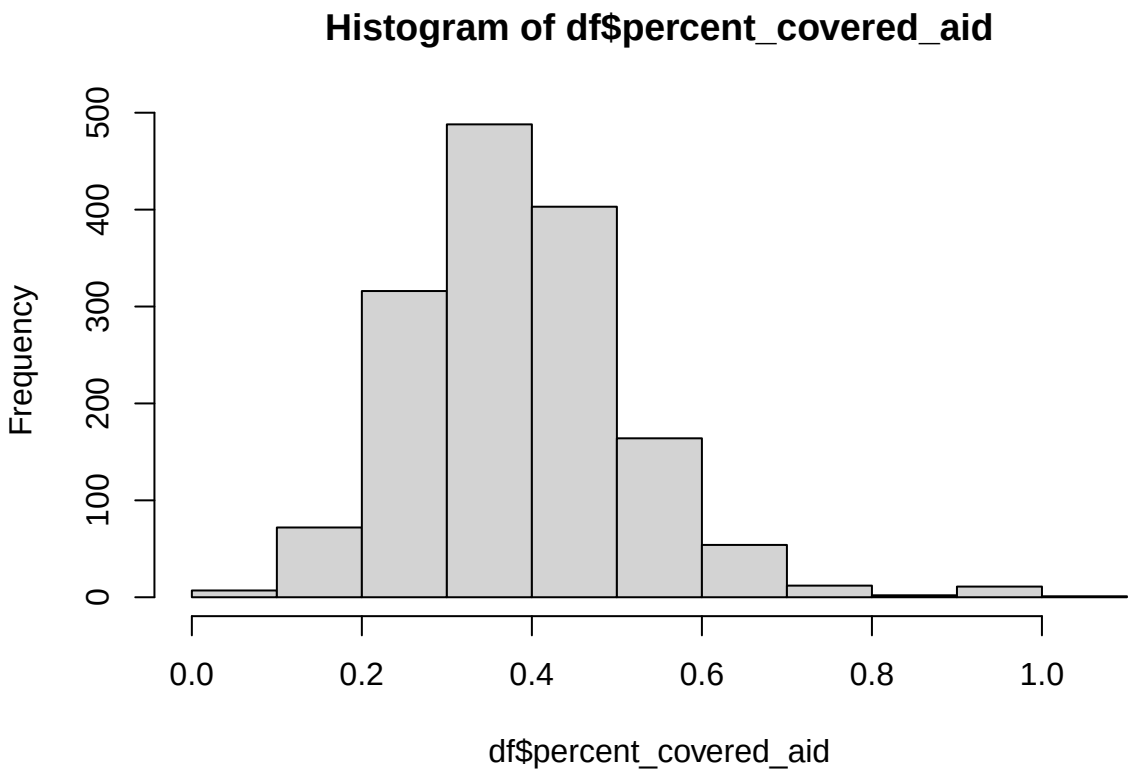



```
df %>%
  ggplot( aes(x = pell_above_avg, y = gradrate_ba_6yrs_total, fill=pell_above_avg)) +
  geom_boxplot() +ggtitle("Comparing Graduation Rates at Institutions where\nLow-Income Students Succee
```

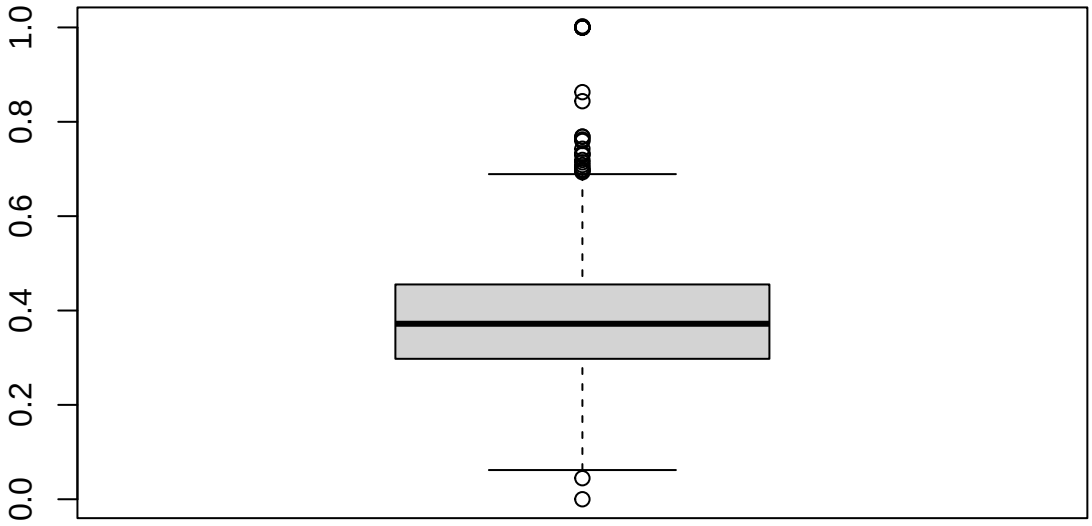


Is a factor of whether Low-Income students succeed percentage of costs covered by grants?

```
# avg amount of federal, state, local, institutional sums up all institutional aid a student can get
# total price for in state tution, off campus. in state vs. out of state typically only counts for publ
df <- df %>% mutate(cost_metric = ifelse(is.na(total_price_for_in_state_students_living_off_campus_not_
                                         total_price_for_in_state_students_living_on_campus,
                                         total_price_for_in_state_students_living_off_campus_not_with_fa

df <- df %>% mutate(percent_covered_aid = average_amount_of_federal_state_local_institutional_or_other_s

hist(df$percent_covered_aid)
```



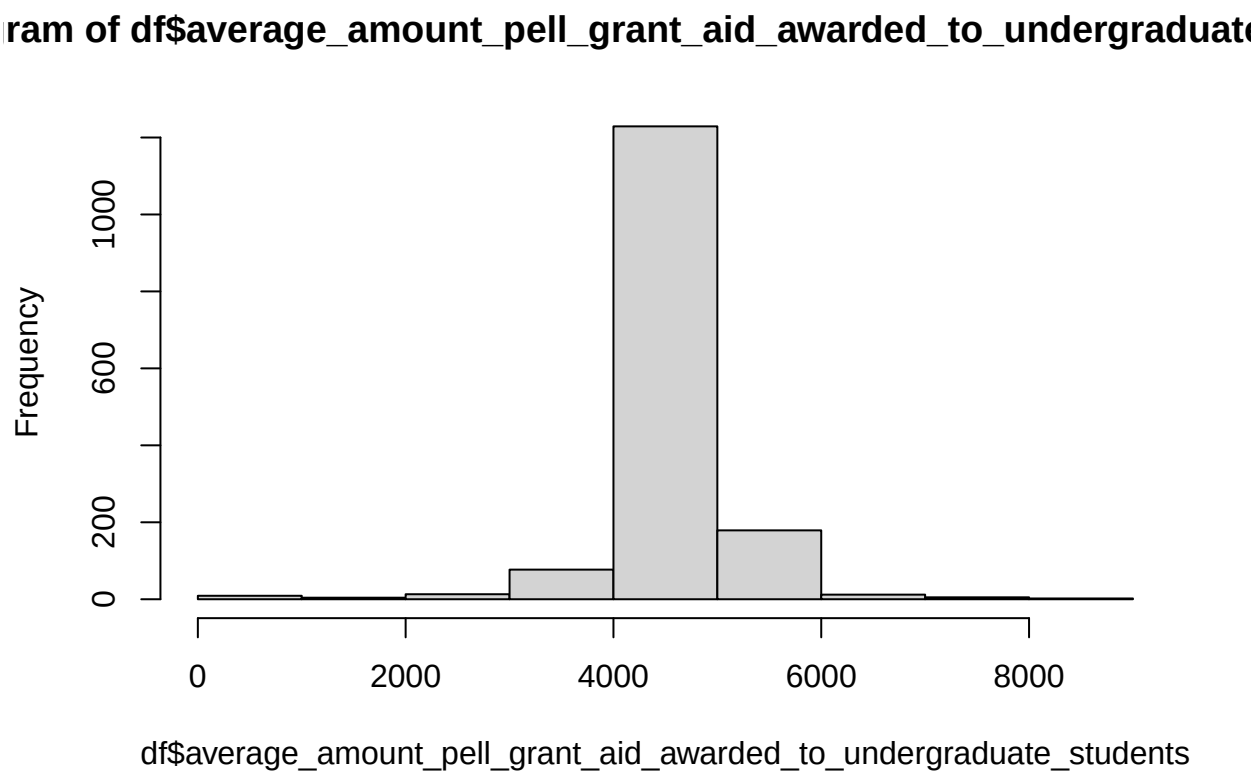

```
boxplot(df$percent_covered_aid)
```



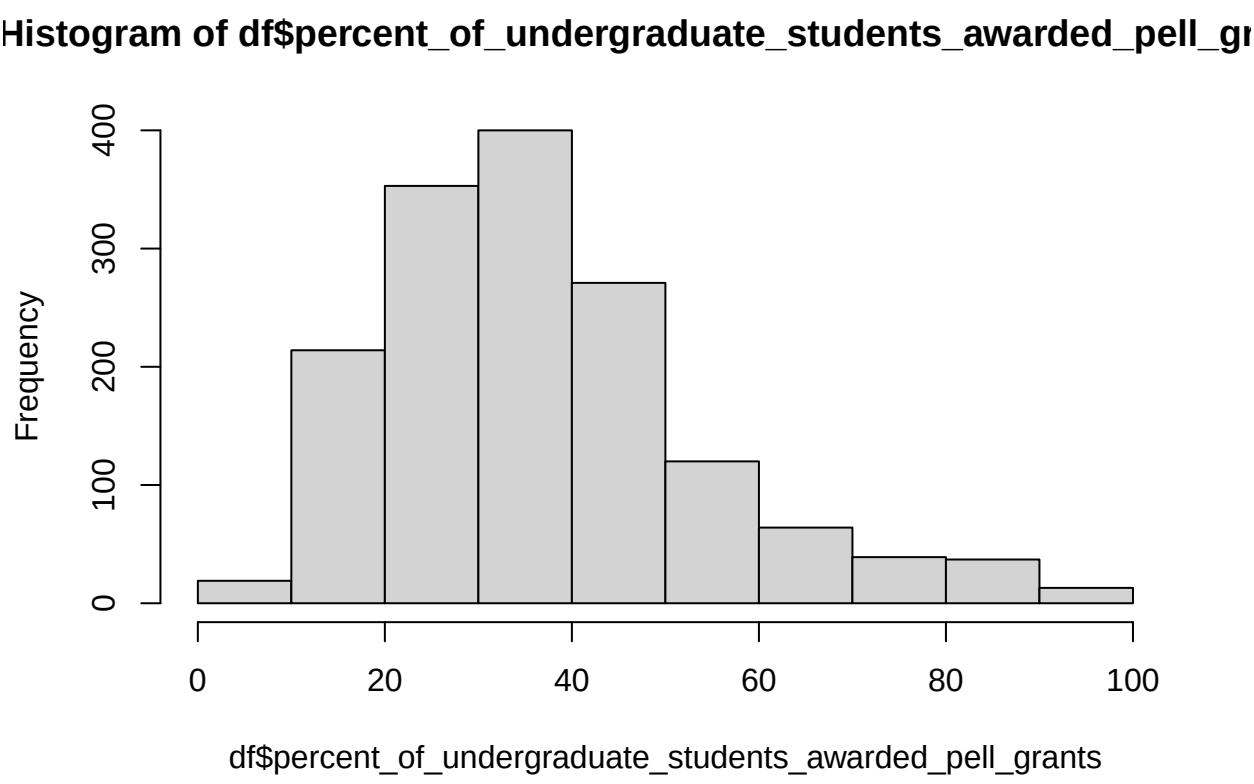

```
pell_df <- df %>% subset(pell_above_avg==TRUE)


nrow(pell_df)
```

```
## [1] 309
```

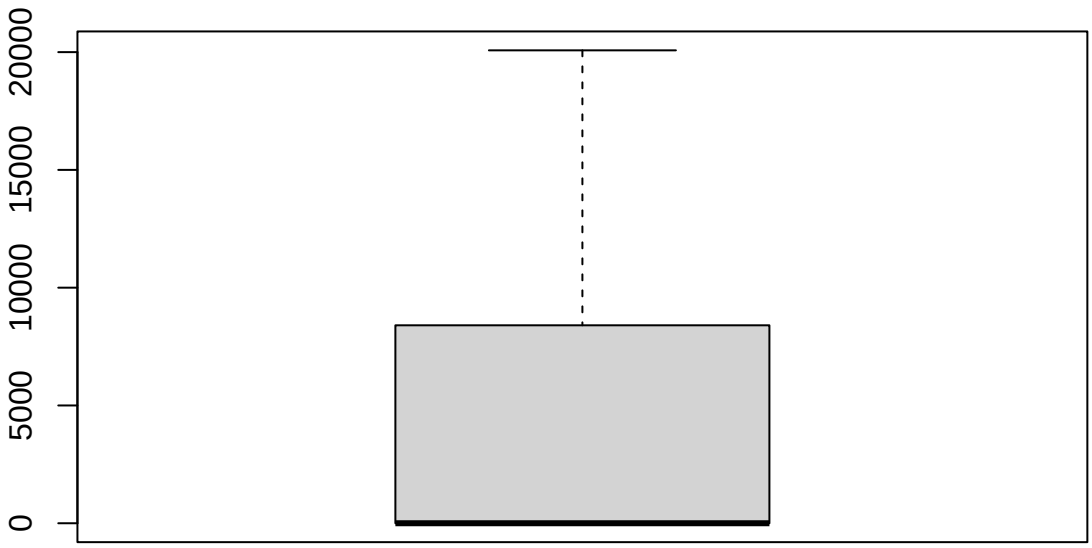
```
hist(df$average_amount_pell_grant_aid_awarded_to_undergraduate_students)
```




```
hist(df$percent_of_undergraduate_students_awarded_pell_grants)
```




```
boxplot(df$average_net_price_income_0_30_000_students_awarded_title_iv_federal_financial_aid)
```



Hypothesis test: Financial Aid Coverage <> Low-Income Students' Grad Rates

```
(mean_aid <- mean(df$percent_covered_aid, na.rm=TRUE))
```

```
## [1] 0.38496
```

```
df <- df %>% mutate(above_avg_aid = ifelse(percent_covered_aid > mean_aid, "above avg. aid", "avg. aid o
print(table(df$above_avg_aid))
```

```
##
##    above avg. aid avg. aid or below
##               714               816
```

```
# get georgetown value
gtown <- df %>% subset(institution_name == 'Georgetown University')
print(gtown$percent_covered_aid)
```

```
## [1] 0.5693391
```

```
# null hypothesis: the average graduation rate for pell students at schools where the average aid given
# alternative hypothesis: the average graduation rate for pell students at schools where the average ai
```

```
t.test(pell_grant_recipients_overall_graduation_rate_within_150_percent_of_normal_time ~ above_avg_aid,
       data = df,
       alternative = "greater")
```

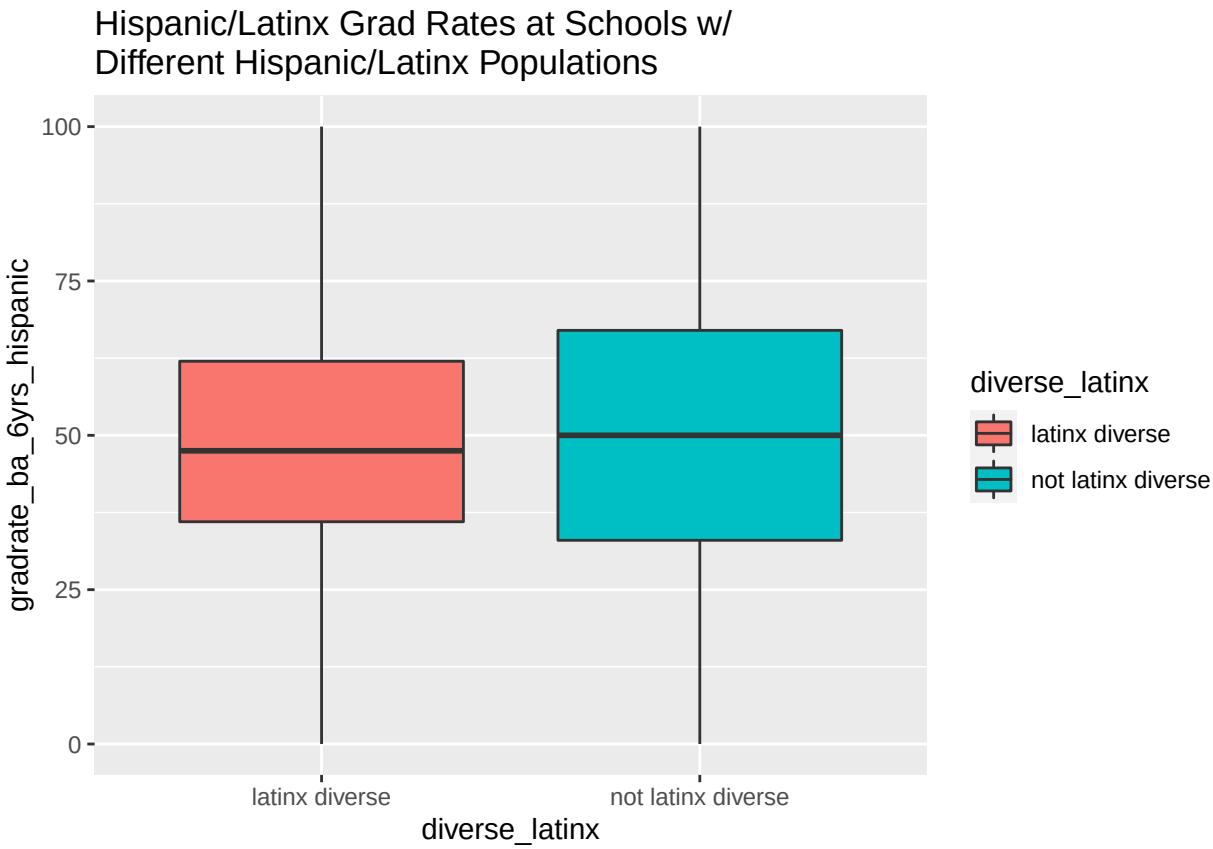
```
##
##  Welch Two Sample t-test
##
## data:  pell_grant_recipients_overall_graduation_rate_within_150_percent_of_normal_time by above_avg_a
## t = 12.605, df = 1410.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group above avg. aid and group avg. aid or b
## 95 percent confidence interval:
##  11.20598      Inf
## sample estimates:
##    mean in group above avg. aid mean in group avg. aid or below
##                         56.37675                         43.48775
```

Hypothesis test: Diversity <> Hispanic/Latinx' Grad Rates

```
(mean_latinx <- mean(df$percent_of_undergraduate_enrollment_that_are_hispanic_latino, na.rm=TRUE))
```

```
## [1] 13.94967
```

```
df <- df %>% mutate(diverse_latinx =  ifelse(percent_of_undergraduate_enrollment_that_are_hispanic_latin
                                             'latinx diverse',
                                             'not latinx diverse'))

df %>%
  ggplot( aes(x = diverse_latinx, y = gradrate_ba_6yrs_hispanic, fill=diverse_latinx)) +
  geom_boxplot() + ggtitle("Hispanic/Latinx Grad Rates at Schools w/\nDifferent Hispanic/Latinx Populat
```




```
# null hypothesis: the average graduation rate for latinx students at schools where the demographic is
# alternative hypothesis: the average graduation rate for latinx students at schools where the demograp
t.test(gradrate_ba_6yrs_hispanic ~ diverse_latinx,
       data = df,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  gradrate_ba_6yrs_hispanic by diverse_latinx
## t = 0.35296, df = 1004.1, p-value = 0.3621
## alternative hypothesis: true difference in means between group latinx diverse and group not latinx di
## 95 percent confidence interval:
##  -1.623404       Inf
## sample estimates:
##     mean in group latinx diverse mean in group not latinx diverse
##                         48.87972                         48.43671
```

Hypothesis test: Diversity <> Black Grad Rates

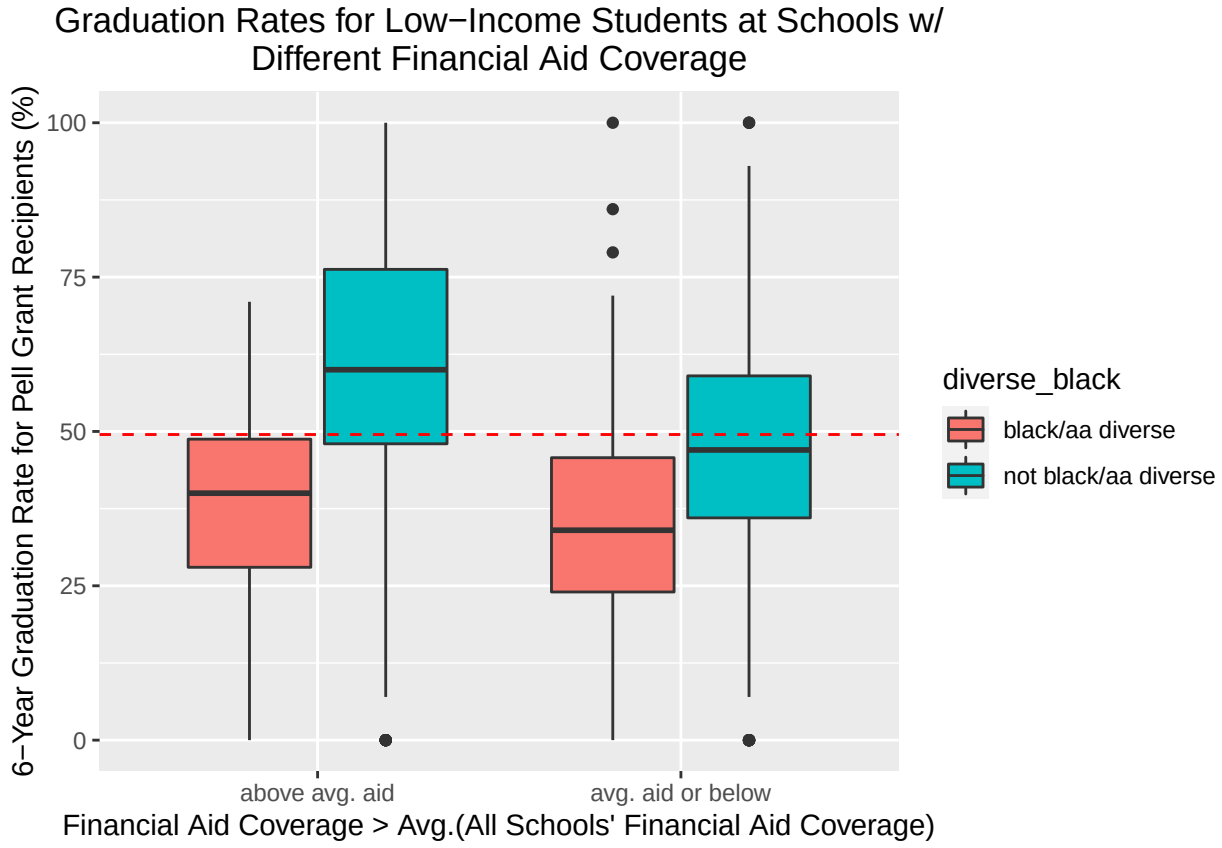
```
(mean_black <- mean(df$percent_of_undergraduate_enrollment_that_are_black_or_african_american, na.rm=TRU
```

```
## [1] 13.11895
```

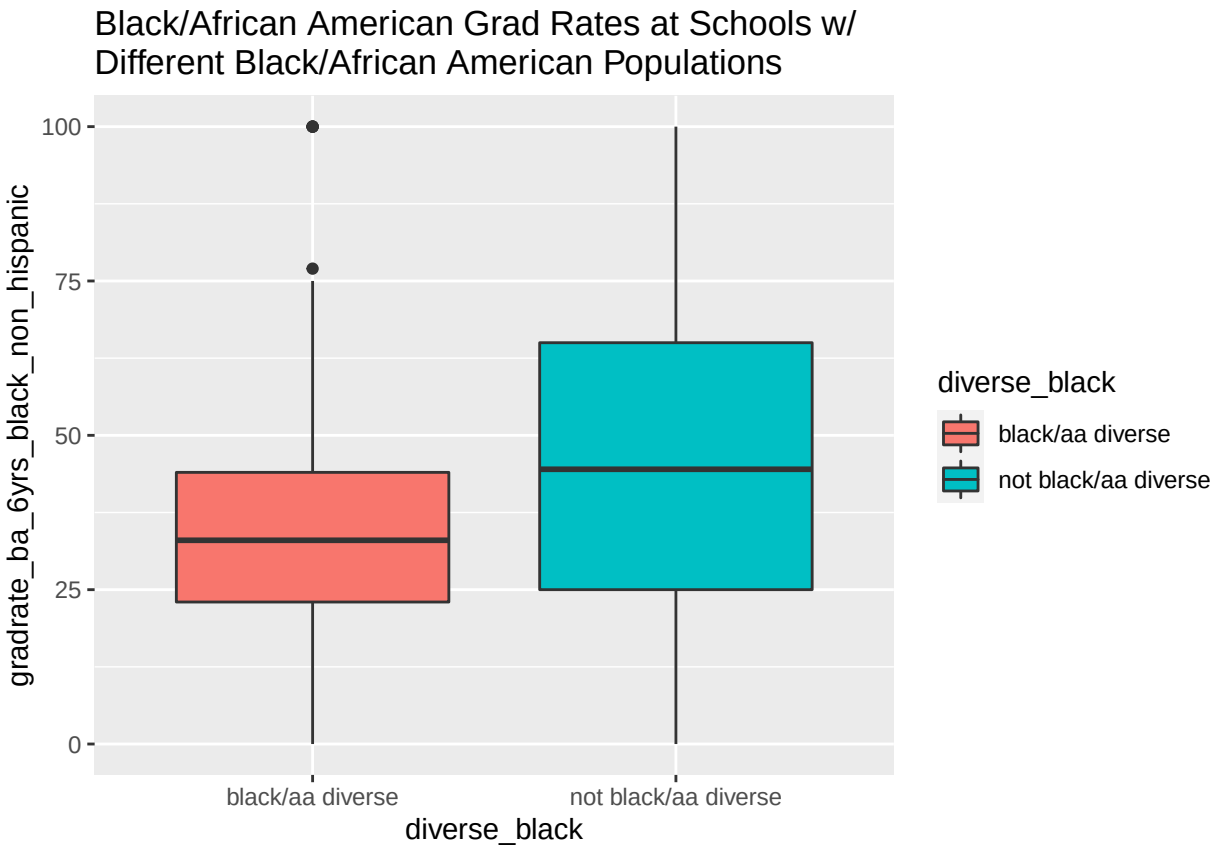
```
df <- df %>% mutate(diverse_black =  ifelse(percent_of_undergraduate_enrollment_that_are_black_or_africa
                                            'black/aa diverse',
                                            'not black/aa diverse'))
print(table(df$diverse_black))
```

```
##
##     black/aa diverse not black/aa diverse
##                  400                 1130
```

```
df %>%
  ggplot( aes(x = above_avg_aid, y = pell_grant_recipients_overall_graduation_rate_within_150_percent_o
  geom_boxplot() + ggtitle("Graduation Rates for Low-Income Students at Schools w/\nDifferent Financial
```




```
df %>%
  ggplot( aes(x = diverse_black, y = gradrate_ba_6yrs_black_non_hispanic, fill=diverse_black)) +
  geom_boxplot() + ggtitle("Black/African American Grad Rates at Schools w/\nDifferent Black/African Am
```




```
# null hypothesis: the average graduation rate for black / african american students at schools where t
# alternative hypothesis: the average graduation rate for black / african american students at schools
t.test(gradrate_ba_6yrs_black_non_hispanic ~ diverse_black,
       data = df,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  gradrate_ba_6yrs_black_non_hispanic by diverse_black
## t = -9.35, df = 1143.1, p-value = 1
## alternative hypothesis: true difference in means between group black/aa diverse and group not black/a
## 95 percent confidence interval:
##  -12.85838       Inf
## sample estimates:
##     mean in group black/aa diverse mean in group not black/aa diverse
##                           33.72500                           44.65841
```

Hypothesis test: Diversity <> Native American Grad Rates

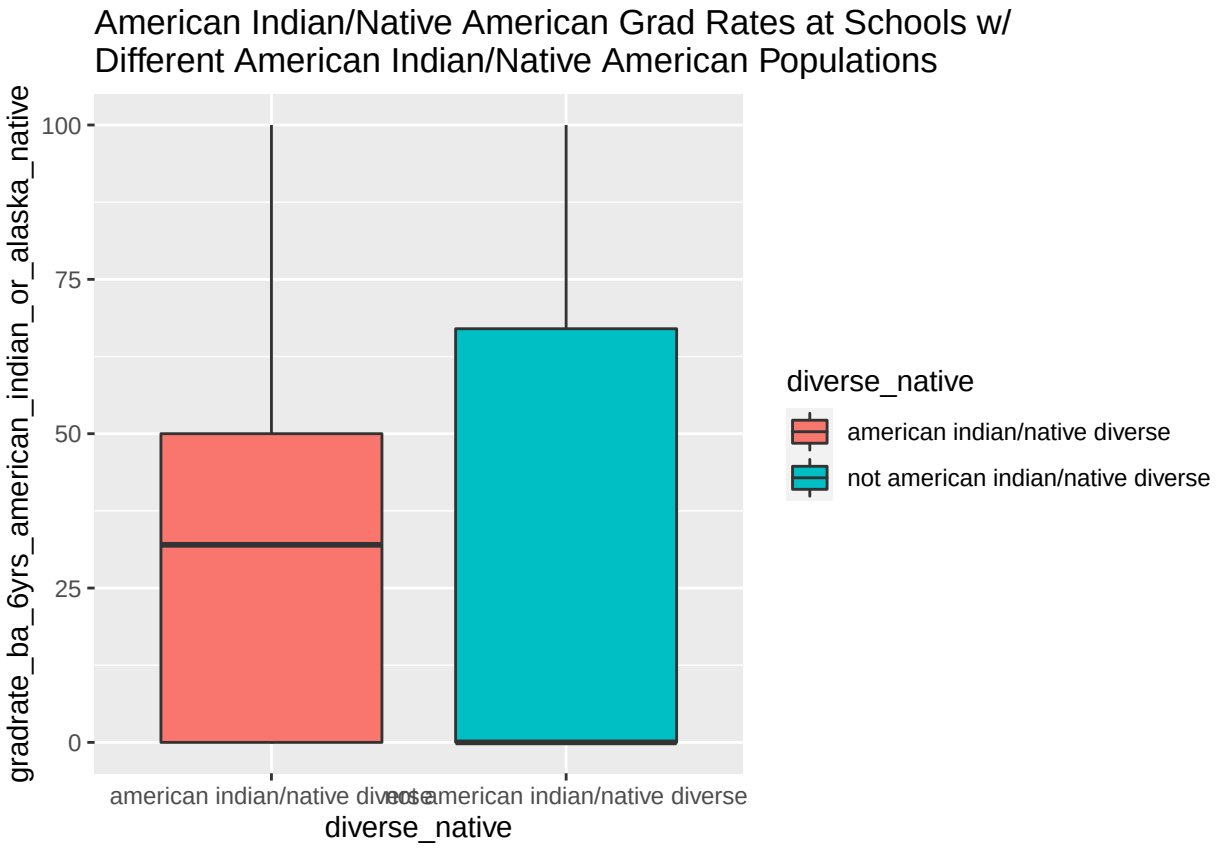
```
(mean_native <- mean(df$percent_of_undergraduate_enrollment_that_are_american_indian_or_alaska_native, r
```

```
## [1] 0.4470588
```

```
df <- df %>% mutate(diverse_native =  ifelse(percent_of_undergraduate_enrollment_that_are_american_indi
                                             'american indian/native diverse',
                                             'not american indian/native diverse'))

print(table(df$diverse_native))
```

```
##
##     american indian/native diverse not american indian/native diverse
##                                347                                1183
```

```
df %>%
  ggplot( aes(x = diverse_native, y = gradrate_ba_6yrs_american_indian_or_alaska_native, fill=diverse_na
  geom_boxplot() + ggtitle("American Indian/Native American Grad Rates at Schools w/\nDifferent Americar
```




```
# null hypothesis: the average graduation rate for american indian/native american students at schools
# alternative hypothesis: the average graduation rate for american indian/native american students at s
t.test(gradrate_ba_6yrs_american_indian_or_alaska_native ~ diverse_native,
       data = df,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  gradrate_ba_6yrs_american_indian_or_alaska_native by diverse_native
```

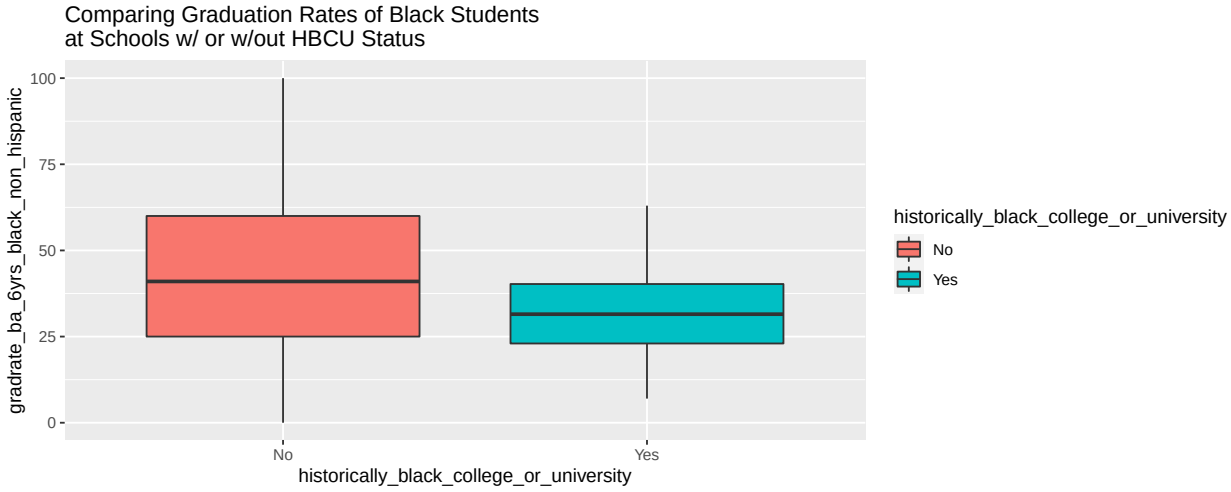
```
## t = 2.0041, df = 646.01, p-value = 0.02274
## alternative hypothesis: true difference in means between group american indian/native diverse and gr
## 95 percent confidence interval:
##  0.76253     Inf
## sample estimates:
##     mean in group american indian/native diverse
##                                         36.29107
## mean in group not american indian/native diverse
##                                         32.00930
```

Hypothesis test: HBCU Status <> Black Grad Rates

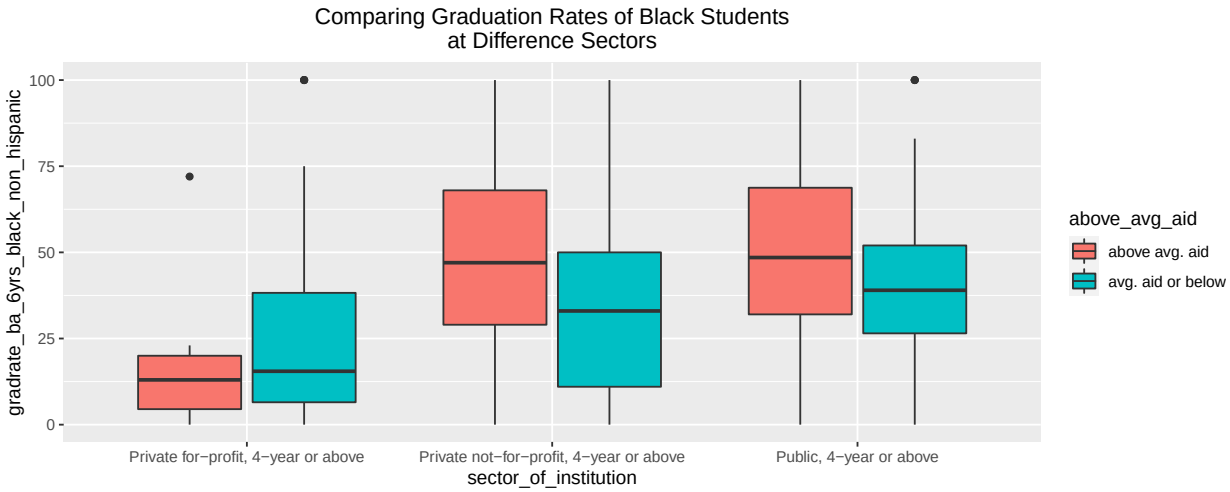
```
print(table(df$historically_black_college_or_university))
```

```
##
##   No  Yes
## 1458   72
```

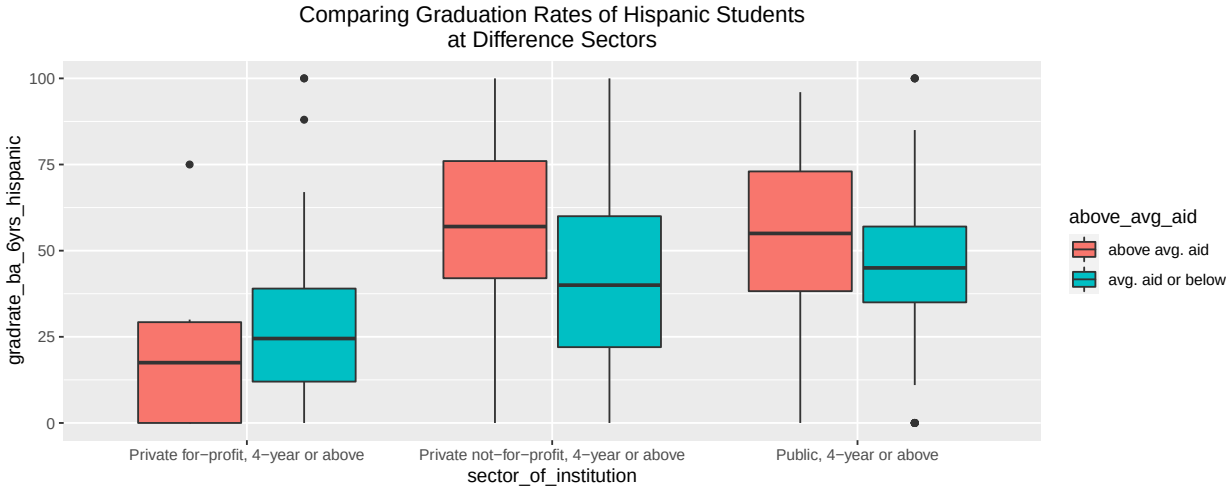
```
df %>%
  ggplot( aes(x = historically_black_college_or_university, y = gradrate_ba_6yrs_black_non_hispanic, fi
  geom_boxplot() + ggtitle("Comparing Graduation Rates of Black Students\nat Schools w/ or w/out HBCU S
```



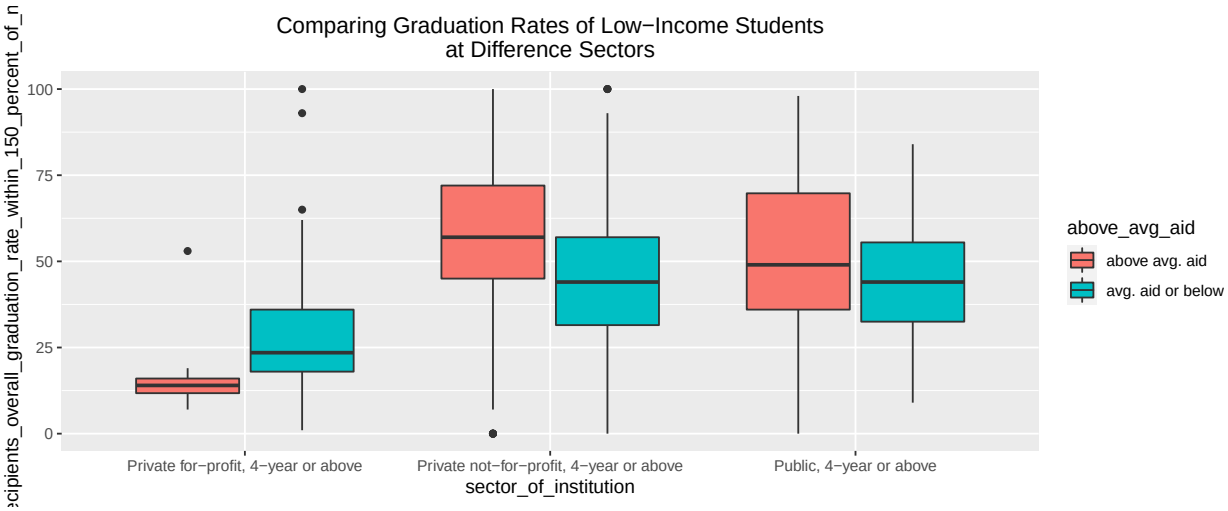

```
df %>%
  ggplot( aes(x = sector_of_institution, y = gradrate_ba_6yrs_black_non_hispanic, fill=above_avg_aid)) +
  geom_boxplot() + ggtitle("Comparing Graduation Rates of Black Students\nat Difference Sectors") + ther
```




```
df %>%
  ggplot( aes(x = sector_of_institution, y = gradrate_ba_6yrs_hispanic, fill=above_avg_aid)) +
  geom_boxplot() + ggtitle("Comparing Graduation Rates of Hispanic Students\nat Difference Sectors") + t
```



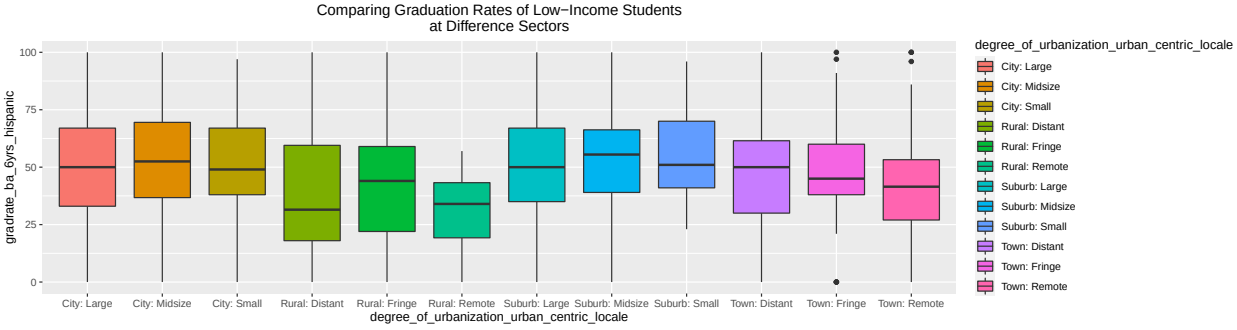

```
df %>%
  ggplot( aes(x = sector_of_institution, y = pell_grant_recipients_overall_graduation_rate_within_150_pe
  geom_boxplot() + ggtitle("Comparing Graduation Rates of Low-Income Students\nat Difference Sectors") +
```




```
# null hypothesis: the average graduation rate for black students at HBCU schools is the same than at n
# alternative hypothesis: the average graduation rate for black students at HBCU schools is statistical
t.test(gradrate_ba_6yrs_black_non_hispanic ~ historically_black_college_or_university,
       data = df,
       alternative = "less")
```

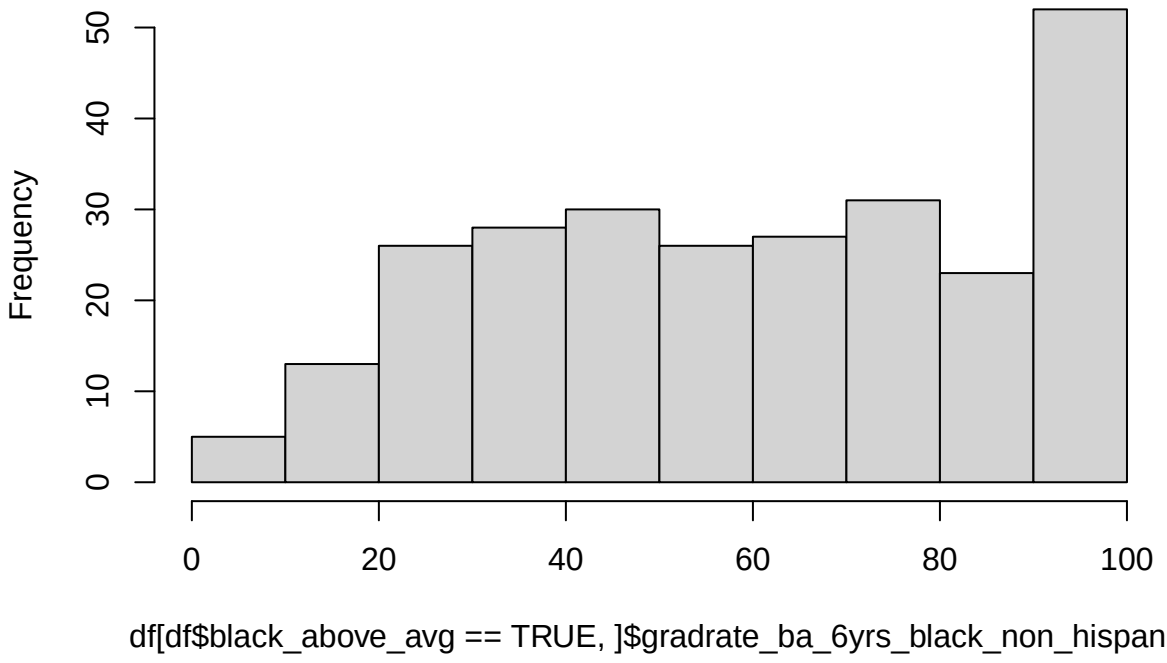
```
##
##  Welch Two Sample t-test
##
## data:  gradrate_ba_6yrs_black_non_hispanic by historically_black_college_or_university
## t = 5.906, df = 104.03, p-value = 1
## alternative hypothesis: true difference in means between group No and group Yes is less than 0
## 95 percent confidence interval:
##      -Inf 12.25896
## sample estimates:
##  mean in group No mean in group Yes
##          42.25034          32.68056
```

```
df %>%
  ggplot( aes(x = degree_of_urbanization_urban_centric_locale, y = gradrate_ba_6yrs_hispanic, fill=degre
  geom_boxplot() + ggtitle("Comparing Graduation Rates of Low-Income Students\nat Difference Sectors") +
```




```
hist(df[df$black_above_avg==TRUE,]$gradrate_ba_6yrs_black_non_hispanic)
```

ram of df[df$black_above_avg == TRUE,]$gradrate_ba_6yrs_black_no




```
temp_df <- df %>% subset(black_above_avg==TRUE)

colnames(temp_df)
```

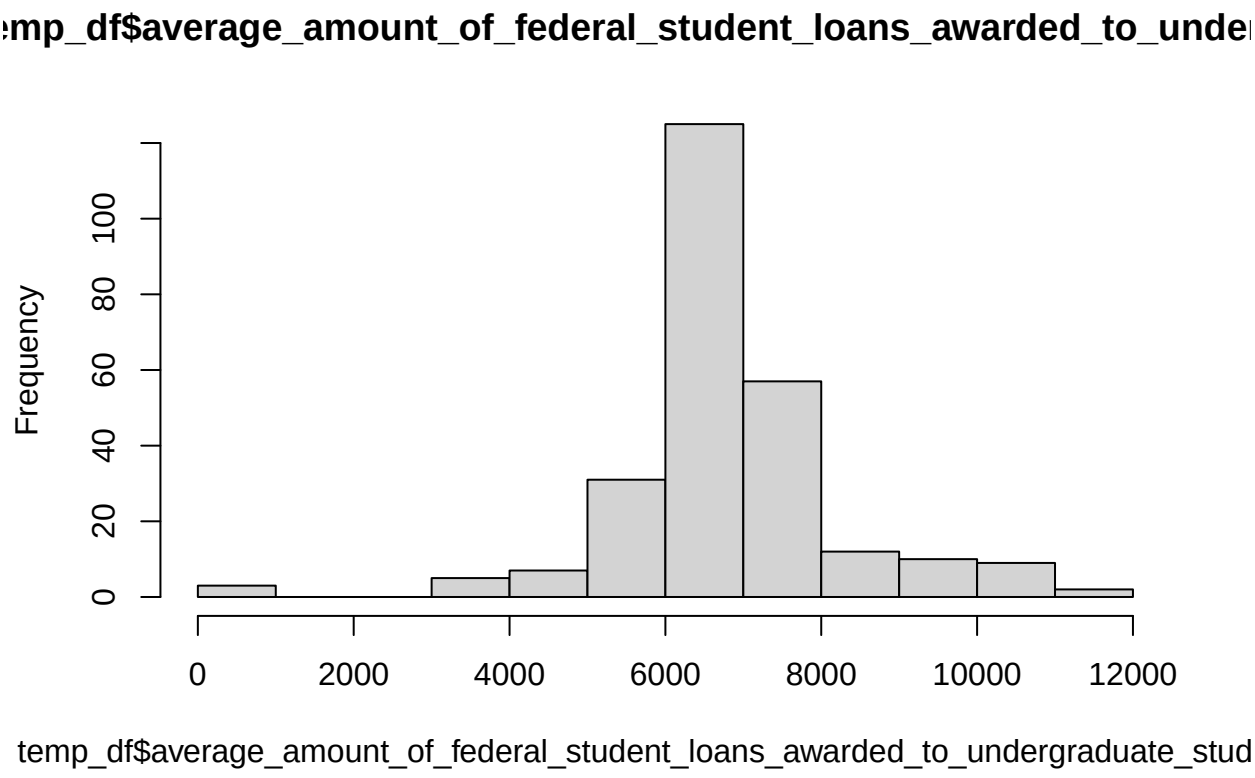
```
##  [1] "unitid"
##  [2] "institution_name"
##  [3] "percent_of_undergraduate_students_awarded_federal_state_local_institutional_or_other_sources_o
##  [4] "average_amount_of_federal_state_local_institutional_or_other_sources_of_grant_aid_awarded_to_u
##  [5] "percent_of_undergraduate_students_awarded_pell_grants"
##  [6] "average_amount_pell_grant_aid_awarded_to_undergraduate_students"
##  [7] "percent_of_undergraduate_students_awarded_federal_student_loans"
##  [8] "average_amount_of_federal_student_loans_awarded_to_undergraduate_students"
##  [9] "average_net_price_students_awarded_grant_or_scholarship_aid"
## [10] "average_net_price_income_0_30_000_students_awarded_title_iv_federal_financial_aid"
## [11] "average_net_price_income_over_110_000_students_awarded_title_iv_federal_financial_aid"
## [12] "published_in_state_tuition_and_fees"
## [13] "published_out_of_state_tuition_and_fees"
## [14] "off_campus_not_with_family_room_and_board"
## [15] "on_campus_room_and_board"
## [16] "total_price_for_in_state_students_living_on_campus"
## [17] "total_price_for_out_of_state_students_living_on_campus"
## [18] "total_price_for_in_state_students_living_off_campus_not_with_family"
## [19] "total_price_for_out_of_state_students_living_off_campus_not_with_family"
## [20] "gradrate_ba_6yrs_total"
## [21] "gradrate_ba_6yrs_men"
```

```
## [22] "gradrate_ba_6yrs_women"
## [23] "gradrate_ba_6yrs_black_non_hispanic"
## [24] "gradrate_ba_6yrs_hispanic"
## [25] "gradrate_ba_6yrs_white_non_hispanic"
## [26] "pell_grant_recipients_overall_graduation_rate_within_150_percent_of_normal_time"
## [27] "subsidized_stafford_loan_recipients_not_receiving_pell_grants_overall_graduation_rate_within_15
## [28] "did_not_receive_pell_grants_or_subsidized_stafford_loans_overall_graduation_rate_within_150_pe
## [29] "historically_black_college_or_university"
## [30] "sector_of_institution"
## [31] "institutional_category"
## [32] "degree_of_urbanization_urban_centric_locale"
## [33] "carnegie_classification_2018_size_and_setting"
## [34] "full_time_retention_rate"
## [35] "part_time_retention_rate"
## [36] "undergraduate_enrollment"
## [37] "percent_of_undergraduate_enrollment_that_are_black_or_african_american"
## [38] "percent_of_undergraduate_enrollment_that_are_hispanic_latino"
## [39] "percent_of_undergraduate_enrollment_that_are_white"
## [40] "percent_of_undergraduate_enrollment_that_are_women"
## [41] "endowment_assets_year_end_per_fte_enrollment"
## [42] "number_of_branches_and_independent_libraries"
## [43] "all_programs_offered_completely_via_distance_education"
## [44] "percent_of_undergraduate_students_enrolled_exclusively_in_distance_education_courses"
## [45] "percent_of_undergraduate_enrollment_that_are_asian"
## [46] "percent_of_undergraduate_enrollment_that_are_american_indian_or_alaska_native"
## [47] "percent_of_undergraduate_enrollment_that_are_native_hawaiian_or_other_pacific_islander"
## [48] "gradrate_ba_6yrs_american_indian_or_alaska_native"
## [49] "gradrate_ba_6yrs_asian"
## [50] "gradrate_ba_6yrs_native_hawaiian_or_other_pacific_islander"
## [51] "endowment_total"
## [52] "finances_spent_research"
## [53] "finances_spent_student_services"
## [54] "finances_spent_public_service"
## [55] "finances_spent_academic_support"
## [56] "finances_spent_instruction"
## [57] "revenue_total"
## [58] "sum_subgroups"
## [59] "sum_race_subgroups"
## [60] "mean_subgroups"
## [61] "diversity_quantiles"
## [62] "pell_above_avg"
## [63] "women_above_avg"
## [64] "black_above_avg"
## [65] "hispanic_above_avg"
## [66] "underrepresented_above_avg"
## [67] "cost_metric"
## [68] "percent_covered_aid"
## [69] "above_avg_aid"
## [70] "diverse_latinx"
## [71] "diverse_black"
## [72] "diverse_native"
```

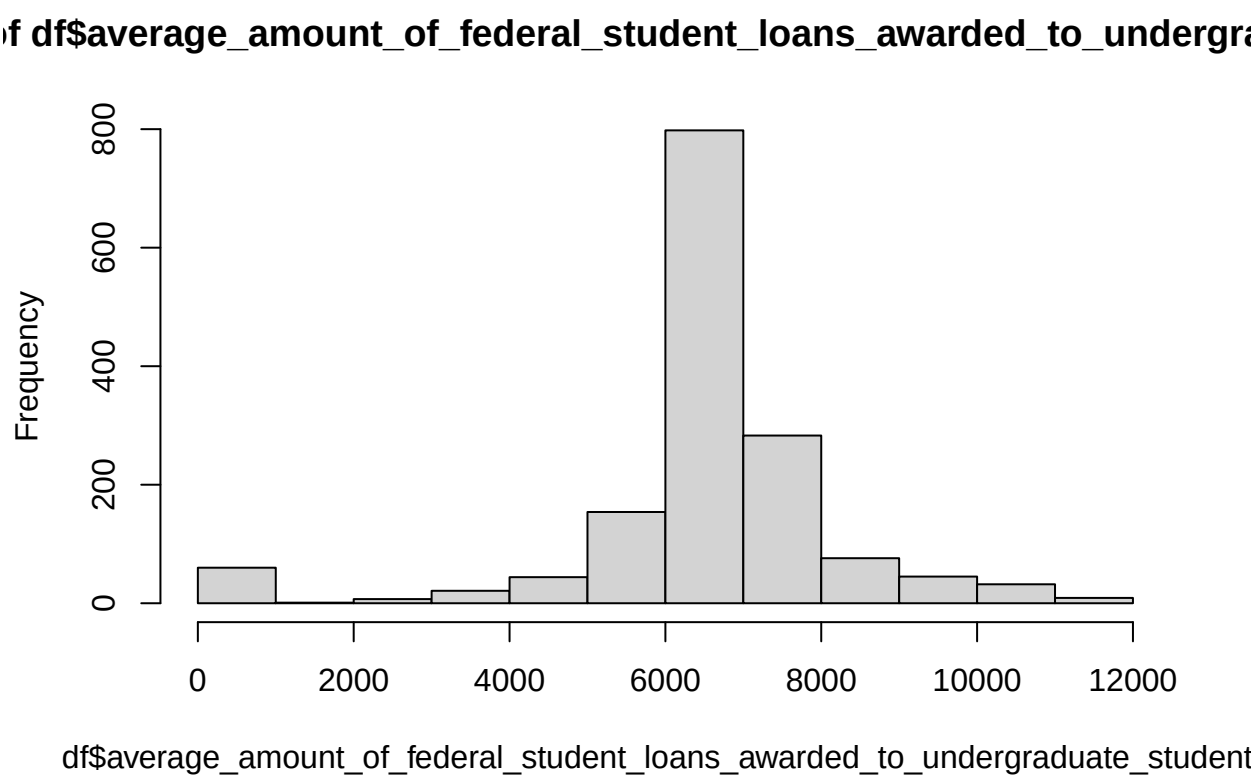
```
print(table(temp_df$sector_of_institution))
```

```
##
##     Private for-profit, 4-year or above Private not-for-profit, 4-year or above
##                                      20                                      140
##                 Public, 4-year or above
##                                     101
```

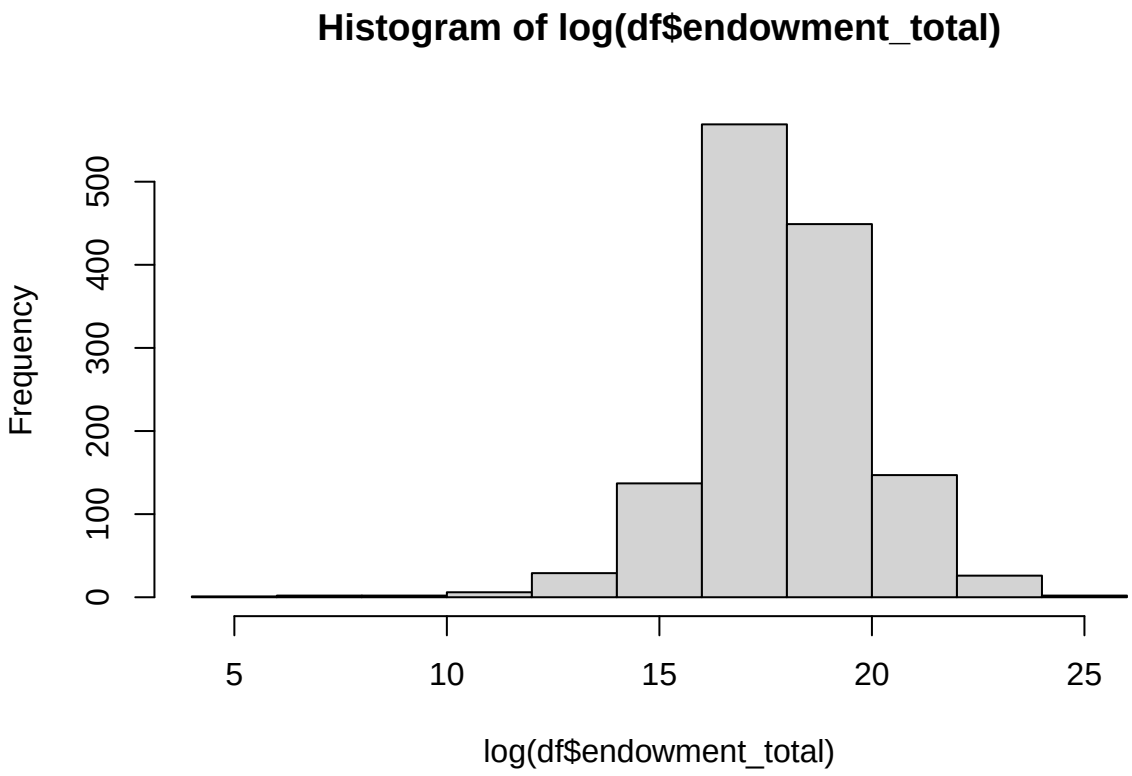
```
hist(temp_df$average_amount_of_federal_student_loans_awarded_to_undergraduate_students)
```




```
hist(df$average_amount_of_federal_student_loans_awarded_to_undergraduate_students)
```




```
hist(log(df$endowment_total))
```




```
df <- df %>% mutate(endowment_quantiles = quantcut(df$mean_subgroups, q = 4, na.rm = TRUE))
df$diversity_quantiles <- mapvalues(df$diversity_quantiles,
                                    from = sort(unique(quantcut(df$mean_subgroups, q = 4, na.rm = TRUE))
                                    to = paste("diversity quantile", 1:4))
```

```
## The following `from` values were not present in `x`: [0.5,9.75], (9.75,13.8], (13.8,19.5], (19.5,50]
```