# DATA 606 - Final Project

Kory Martin

2023-05-13

## Contents



## 1. Introduction

For this project, we will explore data from the California Department of Education in an effort to determine if California charter schools are better than California public schools.

We will attempt to evaluate this question by comparing the mean performance across charter schools and public schools across the following areas:

1. Suspension Rates
2. Graduation Rates
3. Dropout Rates
4. College Going Rates
5. Chronic Absenteeism Rates

For each of these areas, we will use a t-test to test the following hypothesis:

$H_0 : \mu_2 - \mu_1 = 0$

$H_a : \mu_2 - \mu_1 \neq 0$

We will conduct these tests at a 95% level of significance.

Furthermore, once we have determined if there is a statistical signficance in the differences between the groups, we will aware a point to the school-type that has the better performance in this area, using the following convention:

**Suspension Rates** - For suspension rates, we will deem the school type that has the lower average suspension rate as being better

**Graduation Rates** - For graduation rates, we will deem the school type that has the higher average suspension rate as being better

**Dropout Rates** - For dropout rates, we will deem the school type that has the lower average dropout rate as being better

**College Going Rates** - For collage going rates, we will deem the school type that has the higher average college admission rate as being better

**Chronic Absenteeism Rates**- For chronic absenteeism rates, we will deem the school type that has the lower chronic absenteeism rate as being

Finally, the school type that has the most points will be considered the better school type.

A summary of our methodology is as follows:

Methodology:

1. Create dataframe by combining data files from the Department of Education website
2. Combine the various datasets into a singular data file
3. Calculate the average rate across all schools by school type
4. Compare the difference in the average rates to determine if they are statistically significant
5. If the difference is statistically significant, award a point to the school type wth the better average in the respective category
6. Determine which school type has the most point, and conclude that they are the better school-type

# 2. Import Data

We begin by importing the different data files from the California Department of Education website

1. Public Schools and Directory - Downloadable files containing general information about California's public schools and districts.
2. Discipline File - Downloadable data about student discipline and the use of behavioral restraints and seclusion disaggregated by ethnicity, gender, program subgroup, and grade span.
3. Graduation and Dropout File - Downloadable data files of the Four-year Adjusted Cohort Graduation Rate (ACGR) and Outcome data reported by race/ethnicity, program subgroup, and gender.

4. Chronic Absenteeism File - Downloadable data files containing student absenteeism data by race/ethnicity, gender, program subgroup, and grade span. Chronic absenteeism counts, cumulative enrollment, and chronic absenteeism rate data are provided.

5. College-Going Rate for High School Completers File (16-month) - College-Going Rate (CGR) for California high school completers 16 months after high school completion reported by race/ethnicity, student group, and academic year.

```
schools_df_raw <- read_delim('https://www.cde.ca.gov/schooldirectory/report?rid=dl1&tp=txt')
```

```
## Rows: 18305 Columns: 47
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (47): CDSCode, NCESDist, NCESSchool, StatusType, County, District, Schoo...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
susp_df_raw <- read_delim('https://www3.cde.ca.gov/demo-downloads/discipline/suspension22-v2.txt')
```

```
## Rows: 225353 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (20): AcademicYear, AggregateLevel, CountyCode, SchoolCode, CountyName, ...
## dbl  (1): DistrictCode
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
grads_df_raw<- read_delim('https://www3.cde.ca.gov/demo-downloads/acgr/acgr22-v2.txt')
```

```
## Rows: 254175 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (33): AcademicYear, AggregateLevel, CountyCode, SchoolCode, CountyName, ...
## dbl  (1): DistrictCode
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ca_df_raw<- read_delim('https://www3.cde.ca.gov/demo-downloads/attendance/chronicabsenteeism22-v2.txt')
```

```
## Rows: 264937 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (12): Academic Year, Aggregate Level, County Code, School Code, County N...
## dbl  (1): District Code
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
cgr_df_raw<- read_delim('https://www3.cde.ca.gov/demo-downloads/cgr/cgr16mo20.txt')
```

```
## Rows: 436467 Columns: 24
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (23): AcademicYear, AggregateLevel, CountyCode, SchoolCode, CountyName, ...
## dbl  (1): DistrictCode
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Make copies of dataframes

schools_df <- schools_df_raw
susp_df <- susp_df_raw
grads_df <- grads_df_raw
ca_df <- ca_df_raw
cgr_df <- cgr_df_raw
```

# 3. Data Wrangling

For this step, the focus is on preparing the different data frames so that they can be joined together and used downstream for our overall analysis.

## a. Schools Directory

With the school's directory data, we created a sub dataframe that only contains the columns we want. Additionally, we ensured that the cds_cods - the unique identifier for each school - is in the appropriate format for use when joining our other data frames.

```
schools_df <- clean_names(schools_df)

schools_mini <- schools_df %>%
  select(cds_code, status_type, county, district, school, charter, eil_code)

schools_mini <- schools_mini %>%
  mutate(cds_code = str_squish(format(as.numeric(cds_code),scientific=FALSE)))
```

## b. Suspensions

For the suspension data, we focused on creating a sub dataframe that only contains school level data, and shows the suspension data for all students at each school. Additionally, we created a cds_code for each school, by combining the county code, district code, and school code for each school.

```
susp_df <- clean_names(susp_df)

susp_mini <- susp_df %>% select(aggregate_level, county_code, district_code, school_code,
                  reporting_category, suspension_rate_total)
```

```
susp_mini <- susp_mini %>% filter(aggregate_level == 'S',
                     reporting_category == 'TA')

susp_mini <- susp_mini %>%
  mutate(cds_code = paste0(county_code, district_code, school_code)) %>%
  mutate(cds_code = str_squish(format(as.numeric(cds_code), scientific = FALSE)))

susp_mini <- susp_mini %>%
  select(cds_code, suspension_rate_total) %>%
  rename(suspension_rate = suspension_rate_total)
```

### c. Graduates

For the Graduates data, we focused on creating a sub dataframe that only contains school level data, and shows the graduation and dropout data for all students at each school. Additionally, we created a cds_code for each school, by combining the county code, district code, and school code for each school.

```
grads_df <- clean_names(grads_df)

grads_mini <- grads_df %>% select(aggregate_level, county_code, district_code, school_code,
                    charter_school, dass, reporting_category, regular_hs_diploma_graduates_rate,
                    dropout_rate)

grads_mini <- grads_mini %>%
  filter(aggregate_level == 'S',
         dass == 'All',
         reporting_category == 'TA',
         charter_school=='All')

grads_mini <- grads_mini %>%
  mutate(cds_code = paste0(county_code, district_code, school_code)) %>%
  mutate(cds_code = str_squish(format(as.numeric(cds_code),scientific=FALSE)))

grads_mini <- grads_mini %>%
  select(cds_code, regular_hs_diploma_graduates_rate,dropout_rate) %>%
  rename(graduation_rate = regular_hs_diploma_graduates_rate)
```

### d. Chronic Absenteeism

For the Chronic Absenteeism data, we focused on creating a sub dataframe that only contains school level data, and shows the chronic absenteeism data for all students at each school. Additionally, we created a cds_code for each school, by combining the county code, district code, and school code for each school.

```
ca_df <- clean_names(ca_df)

ca_mini <- ca_df %>%
  select(aggregate_level, county_code, district_code, school_code, reporting_category,
         chronic_absenteeism_rate)

ca_mini <- ca_mini %>%
  filter(aggregate_level == 'S', reporting_category == 'TA')
```

```
ca_mini <- ca_mini %>%
  mutate(cds_code = paste0(county_code, district_code, school_code)) %>%
  mutate(cds_code = str_squish(format(as.numeric(cds_code),scientific=FALSE))) %>%
  select(cds_code, chronic_absenteeism_rate)
```

### e. College Going Rate

Finally for the college going rates data, we focused on creating a sub dataframe that only contains school level data, and shows the college going rate data for all students at each school. Additionally, we created a cds_code for each school, by combining the county code, district code, and school code for each school.

```
cgr_df <- clean_names(cgr_df)

cgr_mini <- cgr_df %>%
  select(aggregate_level, county_code, district_code, school_code, reporting_category, school_name,
         completer_type, alternative_school_accountability_status, college_going_rate_total_16_months,
         charter_school)

cgr_mini <- cgr_mini %>% filter(aggregate_level == 'S',
                                reporting_category == 'TA',
                                alternative_school_accountability_status == 'All',
                                completer_type == 'TA',
                                charter_school == 'All')

cgr_mini <- cgr_mini %>%
  mutate(cds_code = paste0(county_code,district_code, school_code)) %>%
  mutate(cds_code = str_squish(format(as.numeric(cds_code),scientific=FALSE)))

cgr_mini <- cgr_mini %>%
  select(cds_code, college_going_rate_total_16_months) %>%
  rename(college_going_rate = college_going_rate_total_16_months)
```

##. f Create joined data frame

Now that we have the necessary data from each file grouped into their separate sub dataframes, we are ready to combined the data into a single data frame that allows us to combine the performance data for each feature of interest with the data for the individual school.

Additionally, I relabeled the data associated with the school level that each school operates. This will make it easier to work with and to remove data for any grades that are not essential for our overall data.

```
merged_df <- left_join(schools_mini,susp_mini) %>%
  left_join(grads_mini) %>%
  left_join(ca_mini) %>%
  left_join(cgr_mini)
```

```
## Joining, by = "cds_code"
## Joining, by = "cds_code"
## Joining, by = "cds_code"
## Joining, by = "cds_code"
```

```
merged_df <- merged_df %>%
  filter(charter != 'No Data')

merged_df <- merged_df %>%
  mutate(eil_code = case_when(
    eil_code == 'A' ~ 'adult',
    eil_code == 'ELEM' ~ 'elementary',
    eil_code == 'ELEMHIGH' ~ 'elementary-high combo',
    eil_code == 'HS' ~ 'high school',
    eil_code == 'INTMIDJR' ~ 'intermediate/middle/junior high',
    eil_code == 'PS' ~ 'preschool',
    eil_code == 'UG' ~ 'ungraded')) %>%
  rename(school_level = eil_code)
```

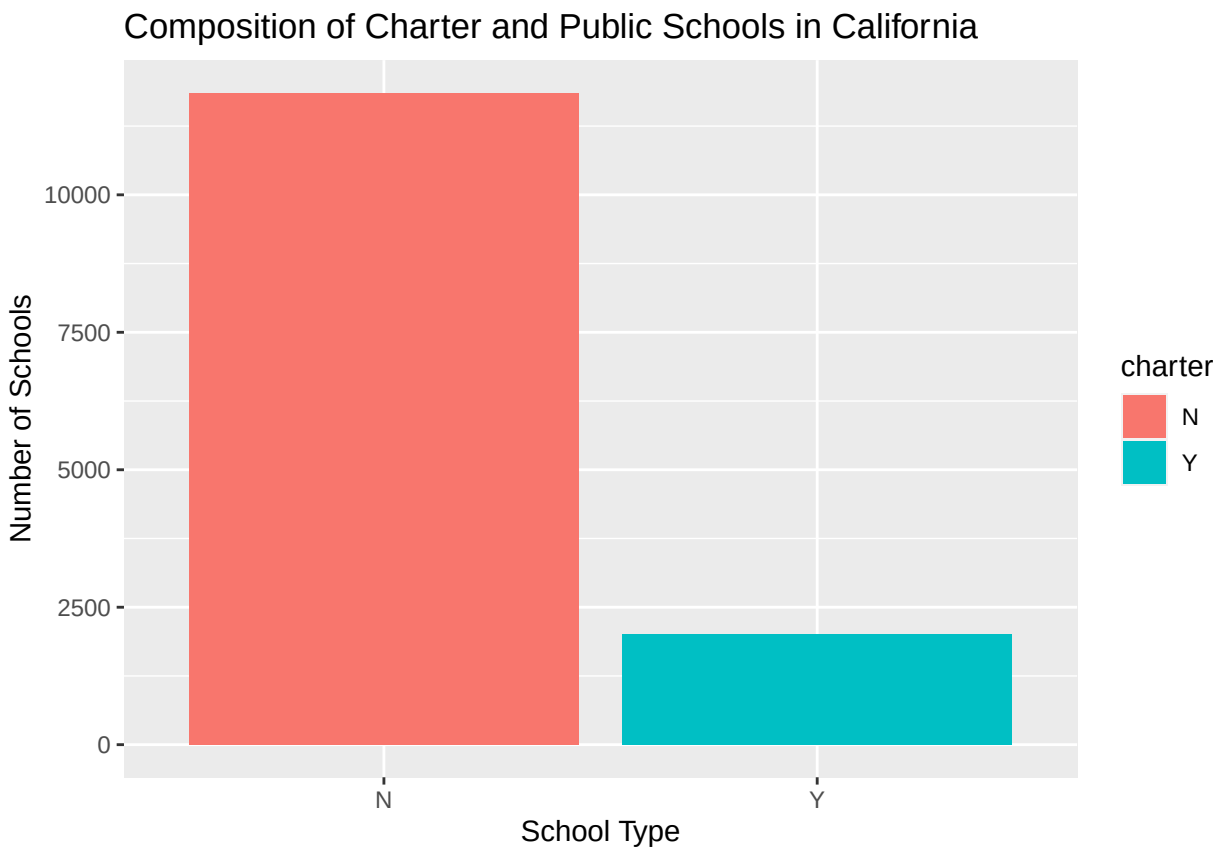
## 4. Data Exploration

I began by exploring the overall merged data, to better understand the available information and to determine how the data is distributed between each school type

```
merged_df <- merged_df %>%
  filter(school_level != 'adult', school_level != 'preschool', school_level != 'ungraded')


ggplot(merged_df) +
  geom_bar(aes(x=charter, fill=charter)) +
  ggtitle("Composition of Charter and Public Schools in California") +
  xlab("School Type") +
  ylab("Number of Schools")
```

Table 1: Number of Schools by Type

| charter | n |
|---|---|
| N | 11857 |
| Y | 2003 |




```
merged_df %>%
  count(charter) %>%
  kable(
    caption = "Number of Schools by Type"
  ) %>%
  kable_material(c("striped"))
```

```
merged_df %>%
  count(charter, school_level) %>%
  spread(key=charter, value=n) %>%
  kable(
    caption = "Number of Schools by Type and School Level"
  ) %>%
  kable_material(c("striped"))
```

```
merged_df %>%
  filter(charter == 'Y') %>%
```

Table 2: Number of Schools by Type and School Level

| school_level | N | Y |
|---|---|---|
| elementary | 6743 | 787 |
| elementary-high combo | 504 | 468 |
| high school | 2931 | 541 |
| intermediate/middle/junior high | 1679 | 207 |

Table 3: Number of Schools by School Level (Charter)

| School Level | N | Pct of Total |
|---|---|---|
| elementary | 787 | 0.39 |
| elementary-high combo | 468 | 0.23 |
| high school | 541 | 0.27 |
| intermediate/middle/junior high | 207 | 0.10 |

```
  group_by(school_level) %>%
  summarize(num_schools = n()) %>%
  mutate(
    pct_schools = round(num_schools/sum(num_schools),2)) %>%
  kable(
    caption = "Number of Schools by School Level (Charter)",
    col.names = c("School Level", "N", "Pct of Total")
  ) %>%
  kable_material(c("striped"))
```

```
merged_df %>%
  filter(charter == 'N') %>%
  group_by(school_level) %>%
  summarize(num_schools = n()) %>%
  mutate(
    pct_schools = round(num_schools/sum(num_schools),2)) %>%
  kable(
    caption = "Number of Schools by School Level (Public)",
    col.names = c("School Level", "N", "Pct of Total")
  ) %>%
  kable_material(c("striped"))
```

```
merged_df <- merged_df %>%
  mutate(suspension_rate = as.double(suspension_rate),
         graduation_rate = as.double(graduation_rate),
```

Table 4: Number of Schools by School Level (Public)

| School Level | N | Pct of Total |
|---|---|---|
| elementary | 6743 | 0.57 |
| elementary-high combo | 504 | 0.04 |
| high school | 2931 | 0.25 |
| intermediate/middle/junior high | 1679 | 0.14 |

```
        dropout_rate = as.double(dropout_rate),
        chronic_absenteeism_rate = as.double(chronic_absenteeism_rate),
        college_going_rate = as.double(college_going_rate))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
summary(merged_df)
```

```
##    cds_code          status_type           county            district
##  Length:13860       Length:13860       Length:13860       Length:13860
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     school            charter           school_level       suspension_rate
##  Length:13860       Length:13860       Length:13860       Min.   : 0.000
##  Class :character   Class :character   Class :character   1st Qu.: 0.200
##  Mode  :character   Mode  :character   Mode  :character   Median : 1.200
##                                                           Mean   : 3.062
##                                                           3rd Qu.: 3.800
##                                                           Max.   :87.500
##                                                           NA's   :3855
##  graduation_rate   dropout_rate     chronic_absenteeism_rate college_going_rate
##  Min.   :  0.00   Min.   :  0.000   Min.   :  0.00           Min.   : 0.00
##  1st Qu.: 77.38   1st Qu.:  1.800   1st Qu.: 20.00           1st Qu.:33.30
##  Median : 91.40   Median :  5.100   Median : 32.00           Median :57.10
##  Mean   : 82.43   Mean   :  9.963   Mean   : 34.09           Mean   :53.89
##  3rd Qu.: 96.00   3rd Qu.: 12.800   3rd Qu.: 45.10           3rd Qu.:73.50
##  Max.   :100.00   Max.   :100.000   Max.   :100.00           Max.   :97.10
##  NA's   :11564    NA's   :11564     NA's   :3955             NA's   :11844
```

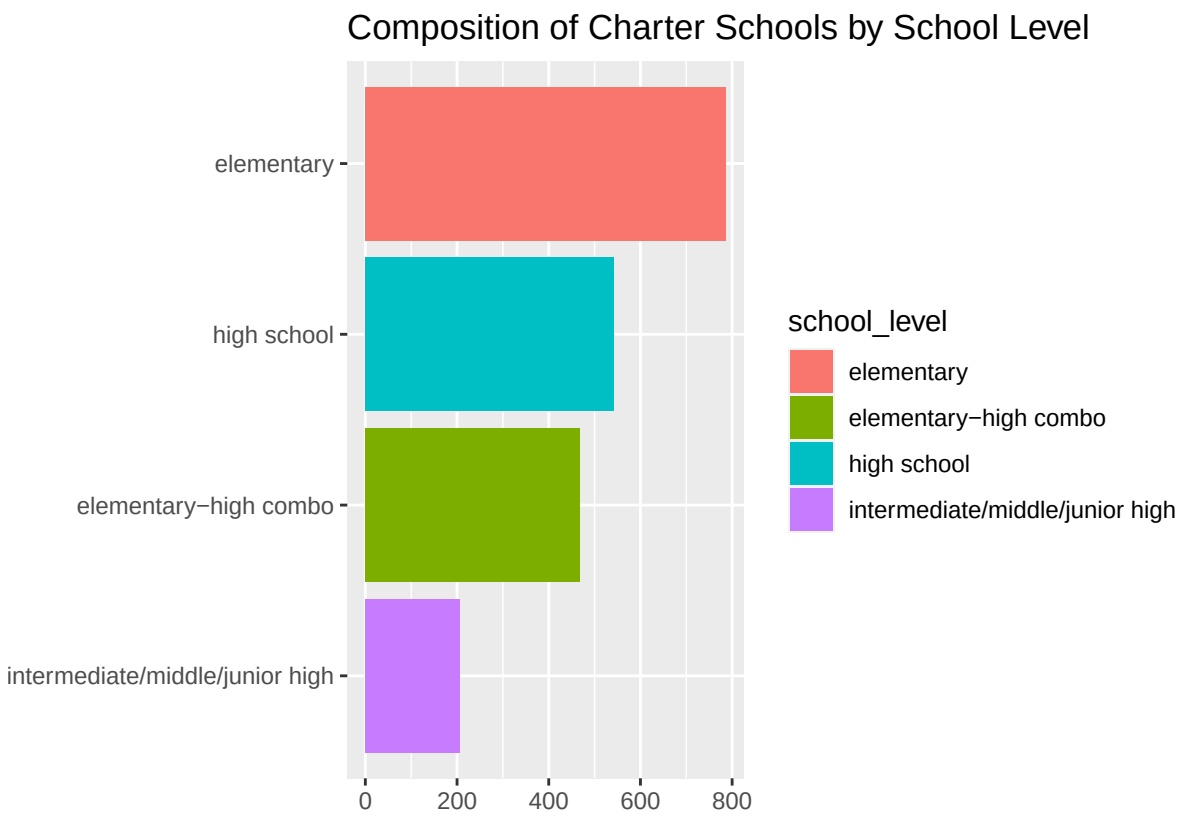
```
merged_df %>%
  group_by(county, charter) %>%
  summarize(num_schools = n_distinct(cds_code)) %>%
  spread(key=charter, value=num_schools) %>%
  mutate(total_schools = N+Y) %>%
  arrange(desc(total_schools)) %>%
  ungroup() %>%
  kable(
    caption = "Number of Schools by County and School Type"
  ) %>%
  kable_material(c("striped"))
```

```
## `summarise()` has grouped output by 'county'. You can override using the
## `.groups` argument.
```

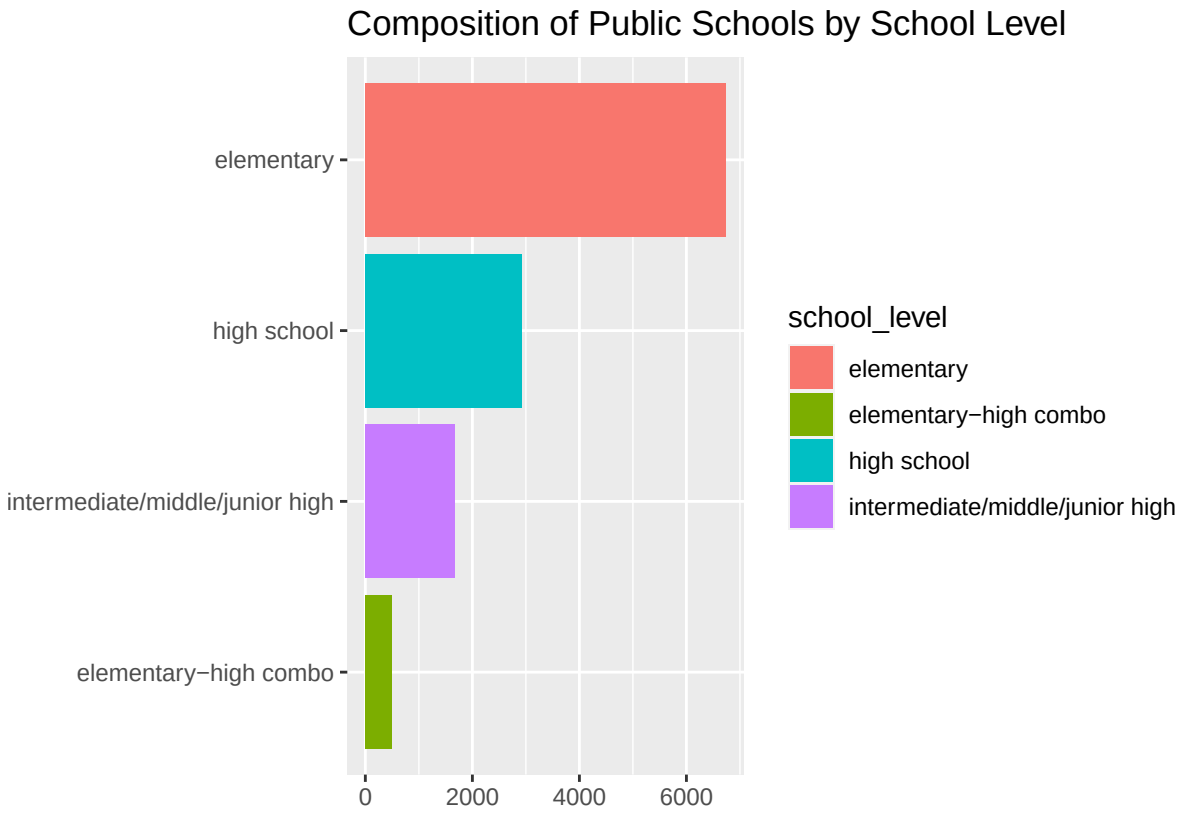
Table 5: Number of Schools by County and School Type

| county | N | Y | total_schools |
|---|---|---|---|
| Los Angeles | 2249 | 536 | 2785 |
| San Diego | 810 | 218 | 1028 |
| San Bernardino | 661 | 87 | 748 |
| Orange | 700 | 43 | 743 |
| Riverside | 616 | 52 | 668 |
| Alameda | 480 | 92 | 572 |
| Santa Clara | 463 | 85 | 548 |
| Sacramento | 459 | 78 | 537 |
| Fresno | 413 | 72 | 485 |
| Kern | 334 | 21 | 355 |
| San Joaquin | 280 | 73 | 353 |
| Contra Costa | 320 | 29 | 349 |
| Sonoma | 186 | 79 | 265 |
| Stanislaus | 220 | 42 | 262 |
| Tulare | 237 | 25 | 262 |
| Ventura | 241 | 18 | 259 |
| San Mateo | 212 | 25 | 237 |
| Monterey | 181 | 18 | 199 |
| Santa Barbara | 154 | 24 | 178 |
| San Francisco | 153 | 23 | 176 |
| Shasta | 138 | 35 | 173 |
| Placer | 139 | 32 | 171 |
| Humboldt | 124 | 21 | 145 |
| Solano | 131 | 14 | 145 |
| Butte | 117 | 24 | 141 |
| Merced | 137 | 4 | 141 |
| Madera | 111 | 11 | 122 |
| San Luis Obispo | 117 | 4 | 121 |
| Santa Cruz | 99 | 17 | 116 |
| Mendocino | 96 | 16 | 112 |
| El Dorado | 81 | 21 | 102 |
| Marin | 91 | 5 | 96 |
| Nevada | 61 | 34 | 95 |
| Yolo | 82 | 11 | 93 |
| Imperial | 82 | 5 | 87 |
| Kings | 68 | 18 | 86 |
| Siskiyou | 78 | 8 | 86 |
| Napa | 67 | 2 | 69 |
| Tehama | 55 | 7 | 62 |
| Sutter | 48 | 13 | 61 |
| Lake | 55 | 4 | 59 |
| Yuba | 45 | 13 | 58 |
| Tuolumne | 49 | 6 | 55 |
| Glenn | 43 | 5 | 48 |
| Lassen | 39 | 9 | 48 |
| Inyo | 38 | 5 | 43 |
| Calaveras | 36 | 1 | 37 |
| Trinity | 32 | 3 | 35 |
| San Benito | 31 | 1 | 32 |
| Mono | 24 | 5 | 29 |
| Plumas | 28 | 1 | 29 |
| Mariposa | 24 | 1 | 25 |
| Modoc | 24 | 1 | 25 |
| Del Norte | 20 | 4 | 24 |
| Amador | 19 | 1 | 20 |

```
merged_df %>%
  filter(charter == 'Y') %>%
  group_by(school_level) %>%
  summarize(n= n()) %>%
  ggplot() +
  geom_bar(aes(x=reorder(school_level, n), y=n, fill=school_level), stat='identity') +
  coord_flip() +
  ggtitle("Composition of Charter Schools by School Level") +
  xlab("") +
  ylab("")
```




```
merged_df %>%
  filter(charter == 'N') %>%
  group_by(school_level) %>%
  summarize(n= n()) %>%
  ggplot() +
  geom_bar(aes(x=reorder(school_level, n), y=n, fill=school_level), stat='identity') +
  coord_flip() +
  ggtitle("Composition of Public Schools by School Level") +
  xlab("") +
  ylab("")
```

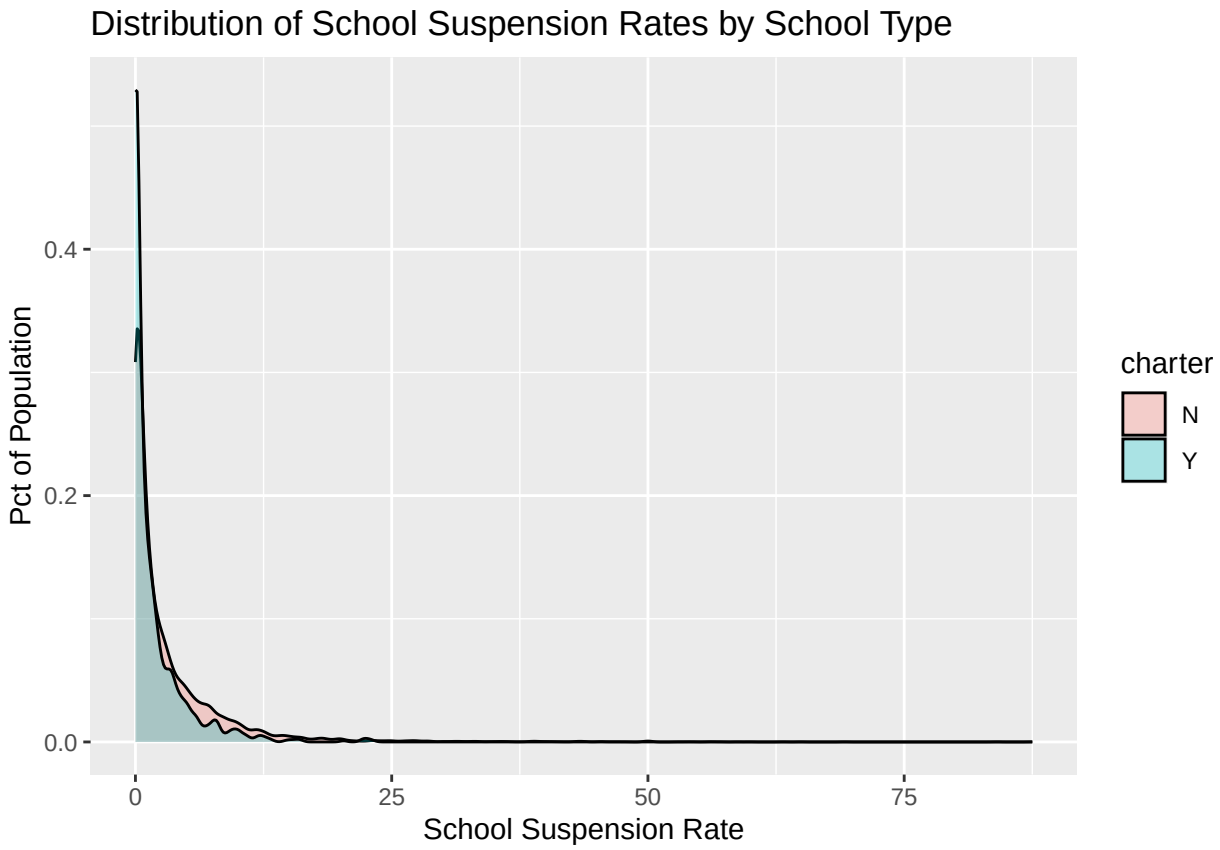



# 5. Data Analysis

## a. Suspension Rate

For suspension rate we will use a t-test to determine if the difference in the average school suspension rate between charter schools and public schools is statistically significant at the 95% level of significance

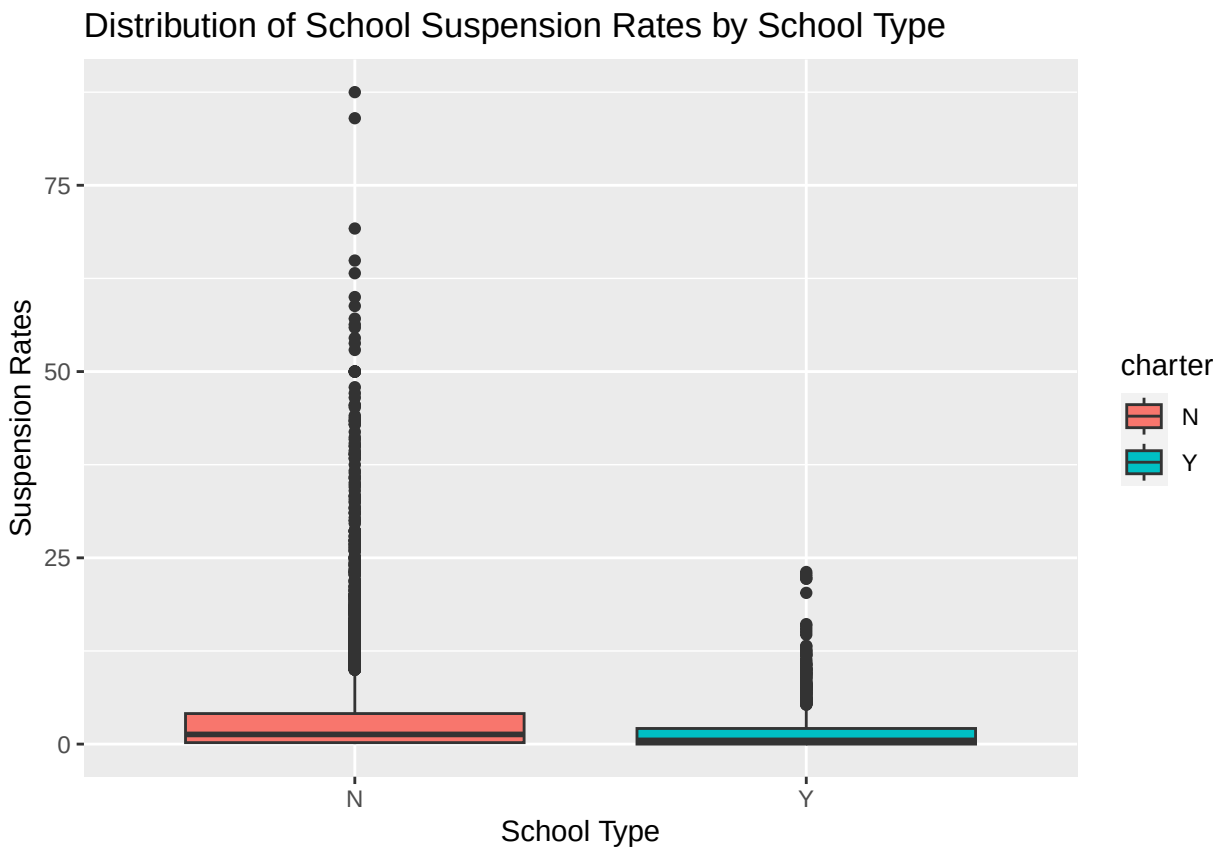
```
m1 <- merged_df %>%
  select(charter, status_type, school_level, suspension_rate) %>%
  filter(!is.na(suspension_rate))


m1_summary <- m1 %>%
  group_by(charter) %>%
  summarize(num_schools = n(),
            avg = round(mean(as.double(suspension_rate)),2),
            sd = round(sd(as.double(suspension_rate)),2))

ggplot(m1,aes(x=suspension_rate,fill=charter)) +
  geom_density(alpha=0.3) +
  ggtitle("Distribution of School Suspension Rates by School Type") +
  xlab("School Suspension Rate") +
  ylab("Pct of Population")
```




```
ggplot(m1) +
  geom_boxplot(aes(x=charter, y=suspension_rate, fill=charter)) +
  ggtitle("Distribution of School Suspension Rates by School Type") +
  xlab("School Type") +
  ylab("Suspension Rates")
```

Table 6: Suspension Rates by School Type

| Charter School | N | Sample Mean | Sample Std. Dev |
|---|---|---|---|
| N | 8720 | 3.26 | 5.61 |
| Y | 1285 | 1.73 | 2.93 |




```
m1_summary %>%
  kable(
    caption = "Suspension Rates by School Type",
    col.names = c("Charter School", "N", "Sample Mean", "Sample Std. Dev")
  ) %>%
  kable_material(c("striped"))
```

```
sr_charter <- (m1 %>% filter(charter == 'Y'))$suspension_rate
sr_public <- (m1 %>% filter(charter == 'N'))$suspension_rate

t.test(x=sr_charter, y=sr_public, alternative="two.sided",conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  sr_charter and sr_public
## t = -15.016, df = 2922.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -1.722124 -1.324312
## sample estimates:
## mean of x mean of y
##  1.734397  3.257615
```

```
suspension_rate = c(1,0)
```

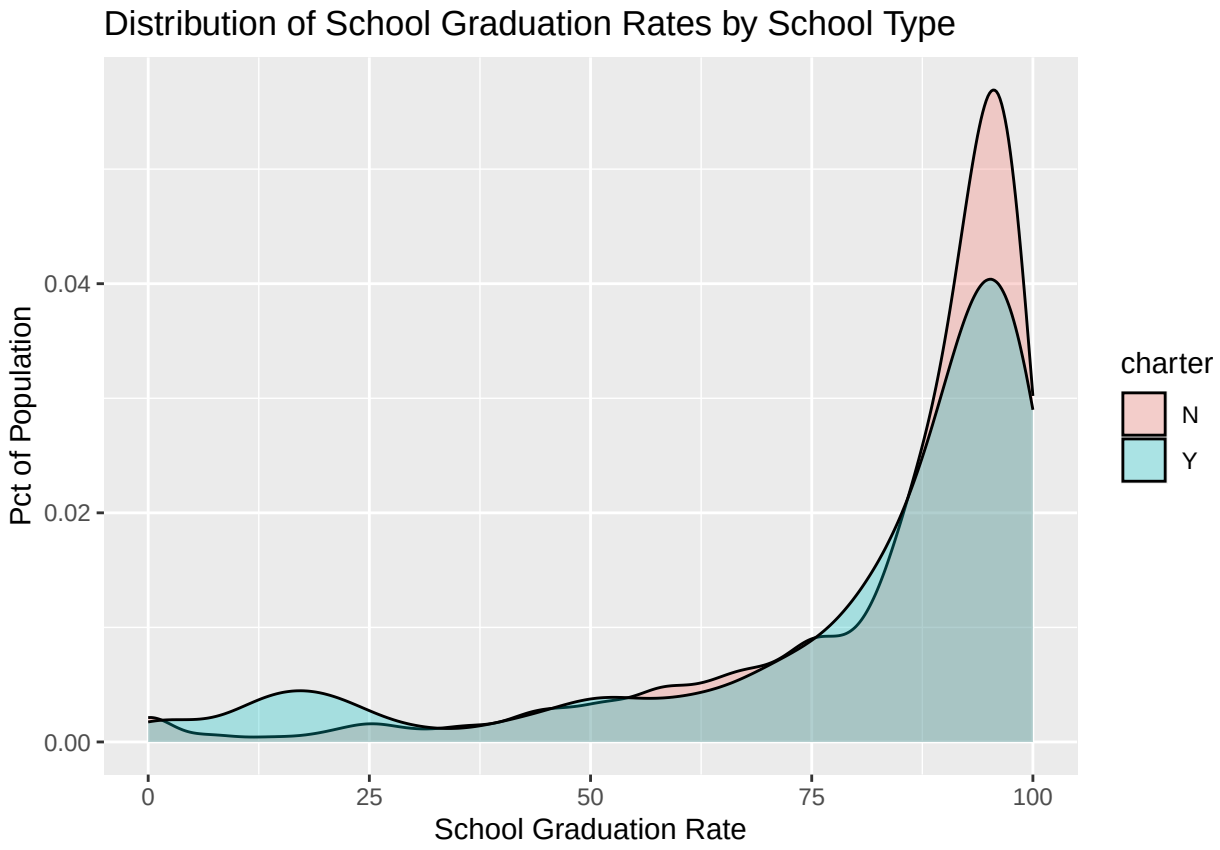
At the 95% level of significance, we have enough evidence to **reject the null hypothesis** that there is no difference in the average suspension rates between charter schools and pubic schools, and therefore we accept the alternative hypothesis that there is a difference in the suspensions rates between charter schools and public schools.

Furthermore, based on our data, we conclude that the suspension rate for charter schools is lower, and therefore better than the suspension rate at public schools. Therefore we conclude that **Charter schools perform better than public schools with respect to suspension rates**.

## b. Graduation Rate

For graduation rate we will use a t-test to determine if the difference in the average school graduation rates between charter schools and public schools is statistically significant at the 95% level of significance

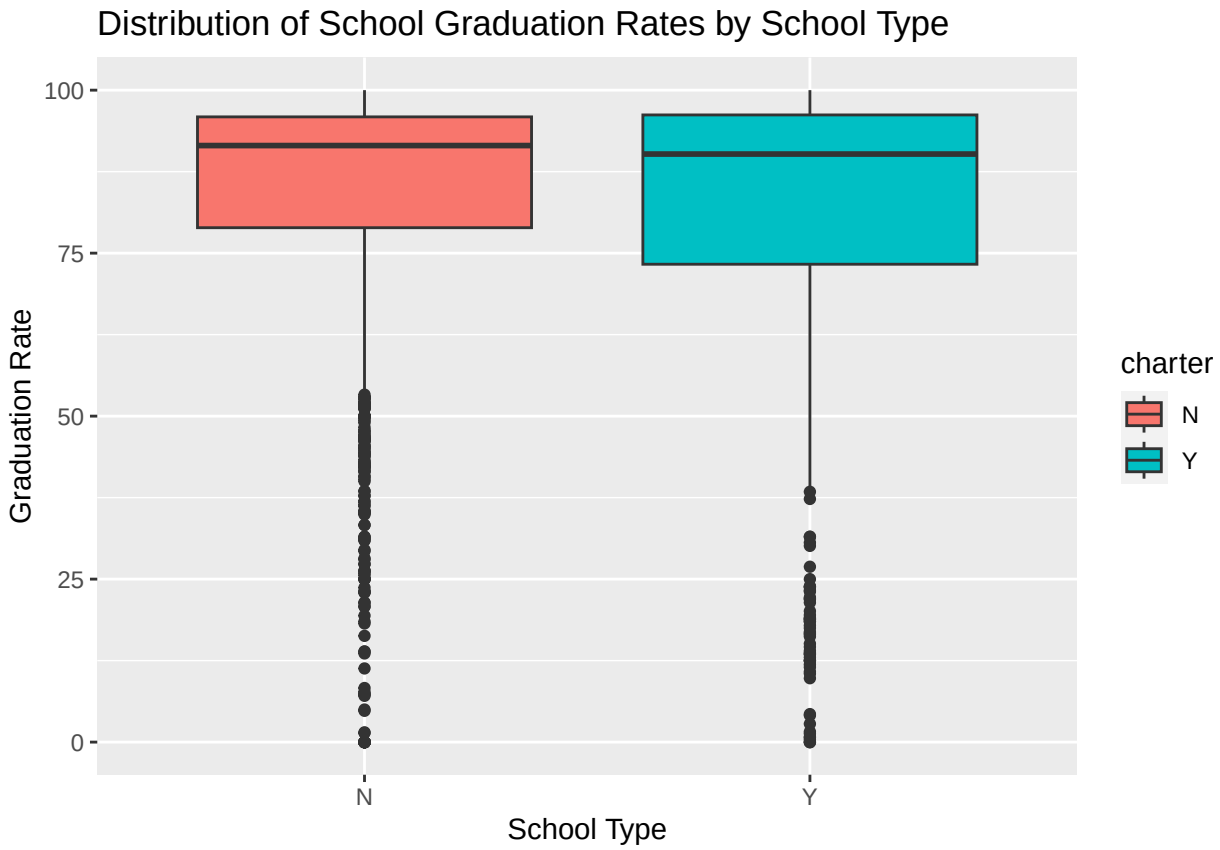
```
m2 <- merged_df %>%
  select(charter, status_type, school_level, graduation_rate)


m2 <- m2 %>%
  filter(school_level != 'adult',
         school_level != 'preschool',
         school_level != 'ungraded')


m2 <- m2 %>%
  filter(!is.na(graduation_rate),
         str_detect(graduation_rate, regex('[[:alnum:]]'))) %>%
  mutate(graduation_rate = as.double(graduation_rate))

m2_summary <- m2 %>%
  group_by(charter) %>%
  summarize(n = n(),
            avg = round(mean(graduation_rate),2),
            sd = round(sd(graduation_rate),2))

ggplot(m2,aes(x=graduation_rate, fill=charter)) +
  geom_density(alpha=0.3) +
  ggtitle("Distribution of School Graduation Rates by School Type") +
  xlab("School Graduation Rate") +
  ylab("Pct of Population")
```




```
ggplot(m2) +
  geom_boxplot(aes(x=charter, y=graduation_rate, fill=charter)) +
  ggtitle("Distribution of School Graduation Rates by School Type") +
  xlab("School Type") +
  ylab("Graduation Rate")
```

Table 7: Graduation Rates by School Type

| Charter School | N | Sample Mean | Sample Std. Dev |
|---|---|---|---|
| N | 1759 | 83.43 | 20.07 |
| Y | 537 | 79.18 | 25.65 |




```
m2_summary %>%
  kable(
    caption = "Graduation Rates by School Type",
    col.names = c("Charter School", "N", "Sample Mean", "Sample Std. Dev")
  ) %>%
  kable_material(c("striped"))
```

```
gr_charter <- (m2 %>% filter(charter == 'Y'))$graduation_rate
gr_public <- (m2 %>% filter(charter == 'N'))$graduation_rate

t.test(x=gr_charter, y=gr_public, alternative="two.sided",conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  gr_charter and gr_public
## t = -3.5241, df = 747.07, p-value = 0.0004507
## alternative hypothesis: true difference in means is not equal to 0
```

| school_level | N | Y |
|---|---|---|
| elementary | 6743 | 787 |
| elementary-high combo | 504 | 468 |
| high school | 2931 | 541 |
| intermediate/middle/junior high | 1679 | 207 |

```
## 95 percent confidence interval:
##  -6.616605 -1.882264
## sample estimates:
## mean of x mean of y
##  79.17598  83.42541
```

```
graduation_rate = c(0,1)
```

At the 95% level of significance, we have enough evidence to **reject the null hypothesis** that there is no difference in the graduation rates between charter schools and pubic schools, and therefore we accept the alternative hypothesis that there is a difference in the average graduation rates between charter schools and public schools.

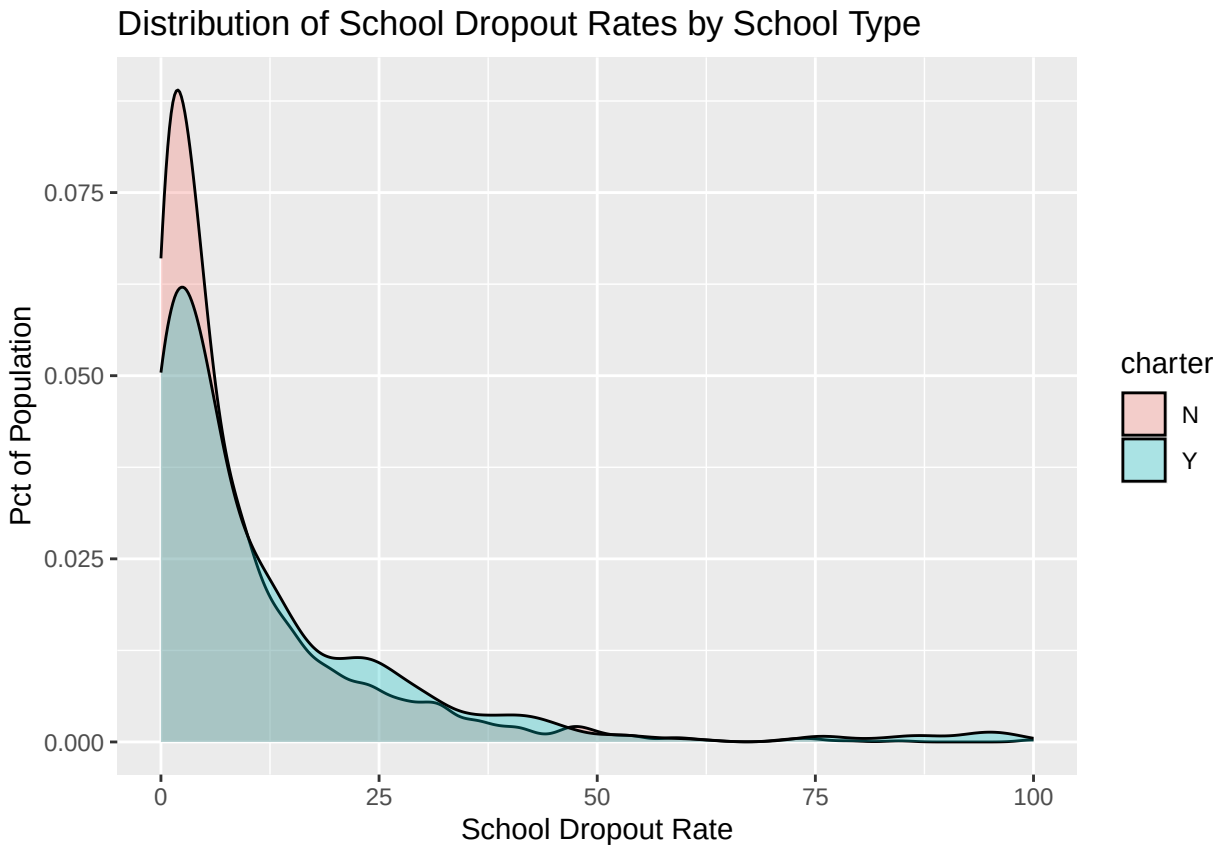
Furthermore, based on our data, we conclude that the graduation rate for charter schools is lower, and therefore worse than than the graduation rate at public schools. Therefore, we conclude that **Public schools perform better than charter schools with respect to graduation rates**.

## c. Dropout Rate

For dropout rate we will use a t-test to determine if the difference in the average school dropout rate between charter schools and public schools is statistically significant at the 95% level of significance

```
m3 <- merged_df %>%
  select(charter, status_type, school_level, dropout_rate)

m3 %>%
  count(charter, school_level) %>%
  spread(key=charter, value = n) %>%
  kable() %>%
  kable_material(c("striped"))
```

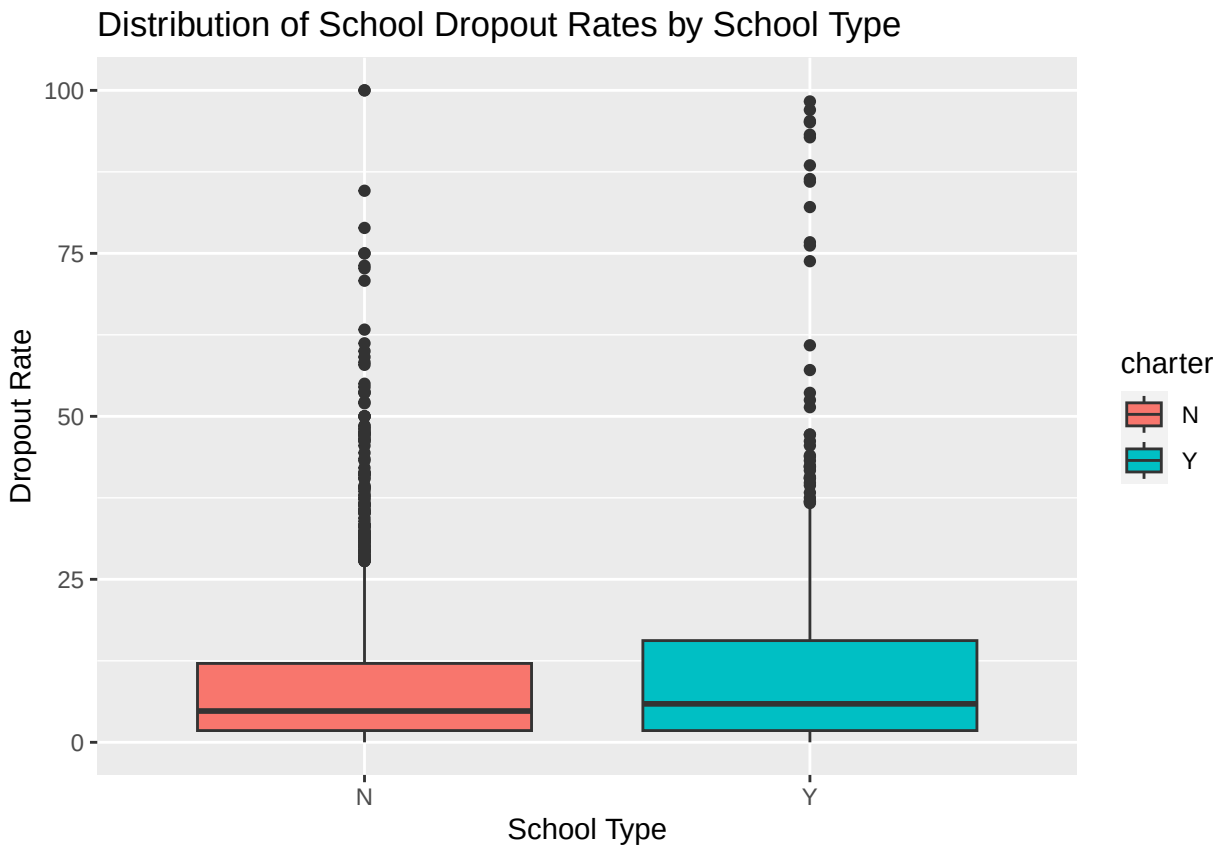
```
m3 <- m3 %>%
  filter(school_level != 'adult',
         school_level != 'preschool',
         school_level != 'ungraded')

m3 <- m3 %>%
  filter(!is.na(dropout_rate),
         str_detect(dropout_rate, regex('[[:alnum:]]'))) %>%
  mutate(dropout_rate = as.double(dropout_rate))

ggplot(m3,aes(x=dropout_rate, fill=charter)) +
  geom_density(alpha=0.3) +
  ggtitle("Distribution of School Dropout Rates by School Type") +
  xlab("School Dropout Rate") +
  ylab("Pct of Population")
```




```
ggplot(m3) +
  geom_boxplot(aes(x=charter, y=dropout_rate, fill=charter)) +
  ggtitle("Distribution of School Dropout Rates by School Type") +
  xlab("School Type") +
  ylab("Dropout Rate")
```

Table 8: Dropout Rates by School Type

| Charter School | N | Sample Mean | Sample Std. Dev |
|---|---|---|---|
| N | 1759 | 9.27 | 11.90 |
| Y | 537 | 12.22 | 16.64 |




```
m3_summary <- m3 %>%
  group_by(charter) %>%
  summarize(n = n(),
            avg = round(mean(dropout_rate),2),
            sd = round(sd(dropout_rate),2))

m3_summary %>%
  kable(
    caption = "Dropout Rates by School Type",
    col.names = c("Charter School", "N", "Sample Mean", "Sample Std. Dev")
  ) %>%
  kable_material(c("striped"))
```

```
dr_charter <- (m3 %>% filter(charter == 'Y'))$dropout_rate
dr_public <- (m3 %>% filter(charter == 'N'))$dropout_rate

t.test(x=dr_charter, y=dr_public, alternative="two.sided",conf.level=0.95)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  dr_charter and dr_public
## t = 3.8232, df = 711.11, p-value = 0.0001433
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.435686 4.466676
## sample estimates:
## mean of x mean of y
## 12.223836  9.272655
```

```
dropout_rate = c(0,1)
```

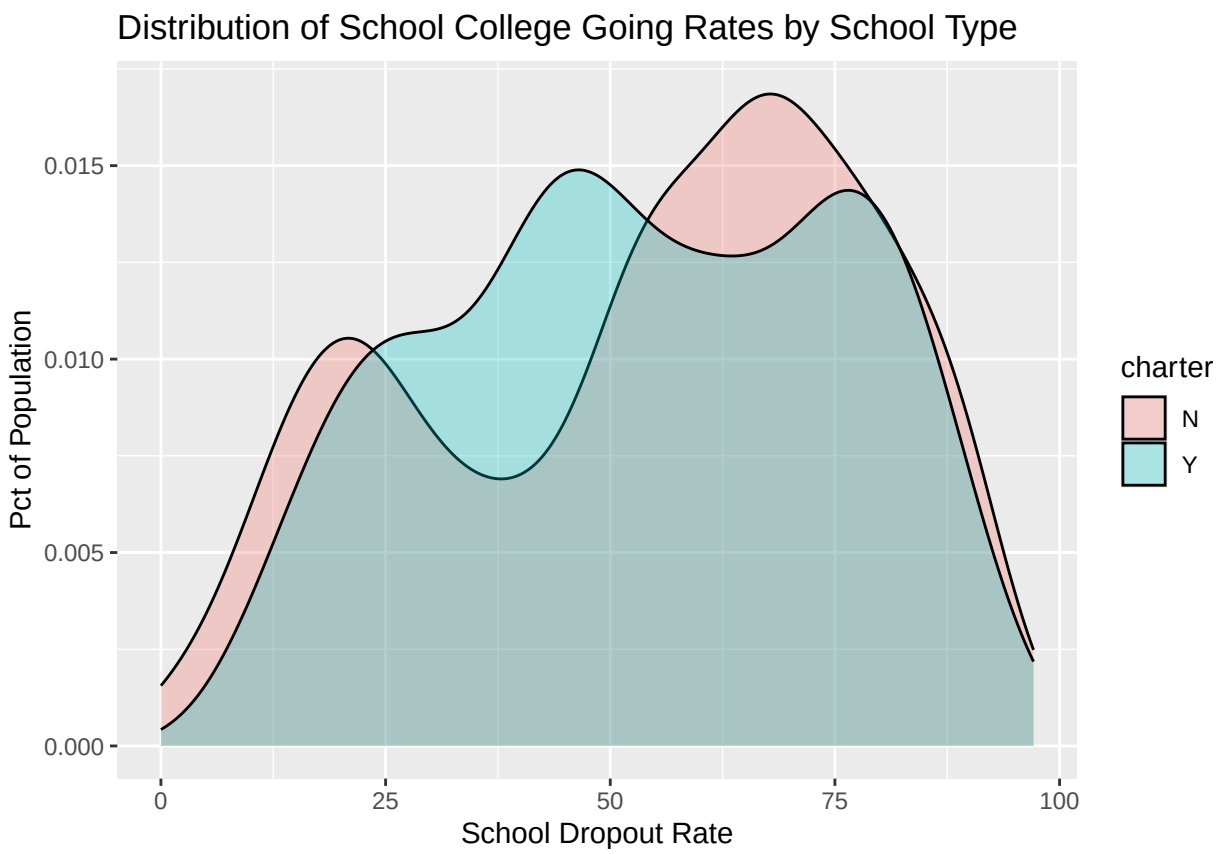
At the 95% level of significance, we have enough evidence to **reject the null hypothesis** that there is no difference in the average dropout rates between charter schools and pubic schools, and therefore we accept the alternative hypothesis that there is a difference in the average dropout rates between charter schools and public schools.

Furthermore, based on our data, we conclude that the dropout rate for public schools is lower, and therefore better than the dropout rate at charter schools. Therefore we conclude that **Public schools perform better than charter schools with respect to dropout rates**.

## d. College Going Rate

For college going rate we will use a t-test to determine if the difference in the average school college going rate between charter schools and public schools (high school and elementary-high combo schools) is statistically significant at the 95% level of significance

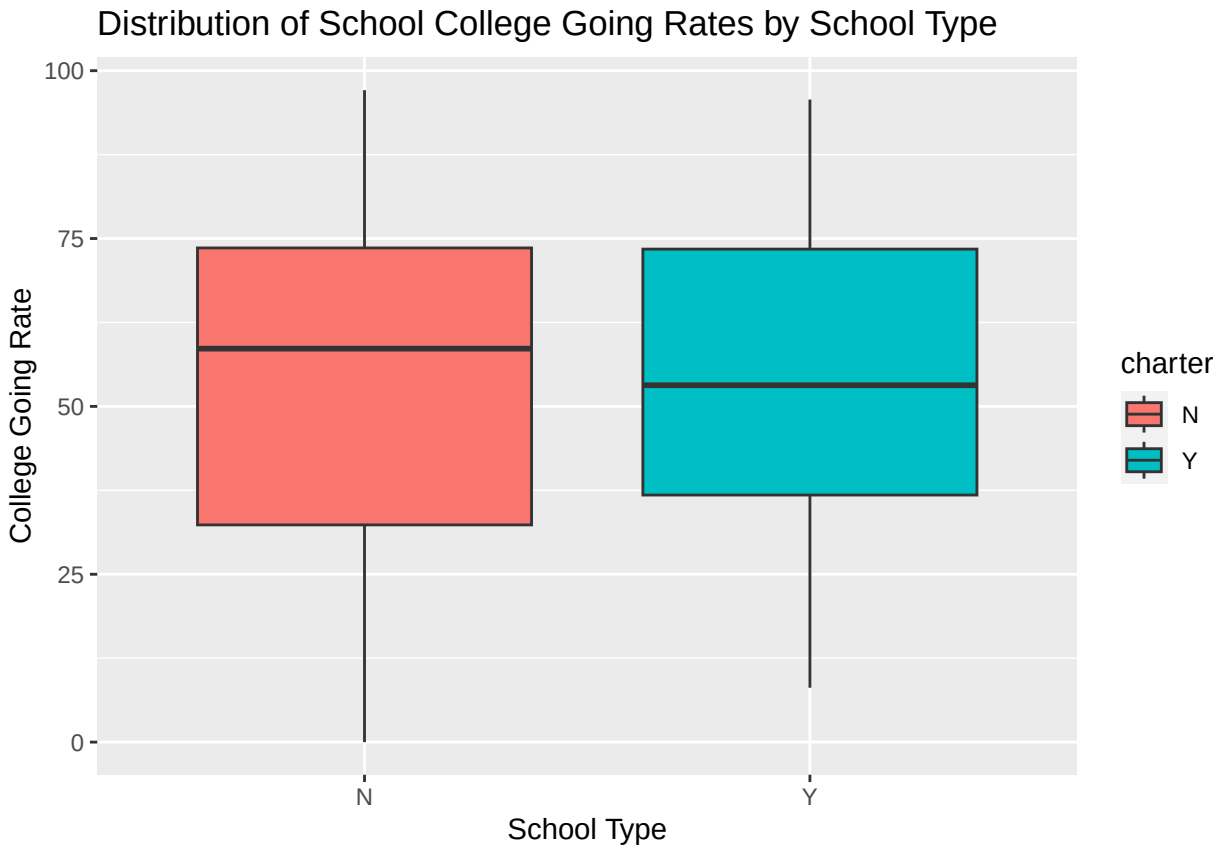
```
m4 <- merged_df %>%
  select(charter, status_type, school_level, college_going_rate)


m4 <- m4 %>%
  filter(school_level %in% c('high school',
                             'elementary-high combo'))

m4 <- m4 %>%
  filter(!is.na(college_going_rate),
         str_detect(college_going_rate, regex('[[:alnum:]]'))) %>%
  mutate(college_going_rate = as.double(college_going_rate))

m4_summary <- m4 %>%
  group_by(charter) %>%
  summarize(n = n(),
            avg = round(mean(college_going_rate),2),
            sd = round(sd(college_going_rate),2))


ggplot(m4,aes(x=college_going_rate, fill=charter)) +
  geom_density(alpha=0.3) +
  ggtitle("Distribution of School College Going Rates by School Type") +
  xlab("School Dropout Rate") +
  ylab("Pct of Population")
```




```
ggplot(m4) +
  geom_boxplot(aes(x=charter, y=college_going_rate, fill=charter)) +
  ggtitle("Distribution of School College Going Rates by School Type") +
  xlab("School Type") +
  ylab("College Going Rate")
```

Table 9: College Going Rates by School Type

| Charter School | N | Sample Mean | Sample Std. Dev |
|---|---|---|---|
| N | 1535 | 54.02 | 24.48 |
| Y | 480 | 53.47 | 22.28 |




```
m4_summary %>%
  kable(
    caption = "College Going Rates by School Type",
    col.names = c("Charter School", "N", "Sample Mean", "Sample Std. Dev")
  ) %>%
  kable_material(c("striped"))
```

```
cgr_charter <- (m4 %>% filter(charter == 'Y'))$college_going_rate
cgr_public <- (m4 %>% filter(charter == 'N'))$college_going_rate

t.test(x=cgr_charter, y=cgr_public, alternative="two.sided",conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  cgr_charter and cgr_public
## t = -0.45707, df = 870.08, p-value = 0.6477
## alternative hypothesis: true difference in means is not equal to 0
```

| school_level | N | Y |
|---|---|---|
| elementary | 6743 | 787 |
| elementary-high combo | 504 | 468 |
| high school | 2931 | 541 |
| intermediate/middle/junior high | 1679 | 207 |

```
## 95 percent confidence interval:
##  -2.888105  1.797037
## sample estimates:
## mean of x mean of y
##  53.47479  54.02033
```

```
college_going_rate = c(0,0)
```

At the 95% level of significance, we are **unable to reject the null hypothesis**, and therefore we conclude that **there is no difference between Charter schools and Public Schools with respect to college going rates**

## e. Chronic Absenteeism Rate

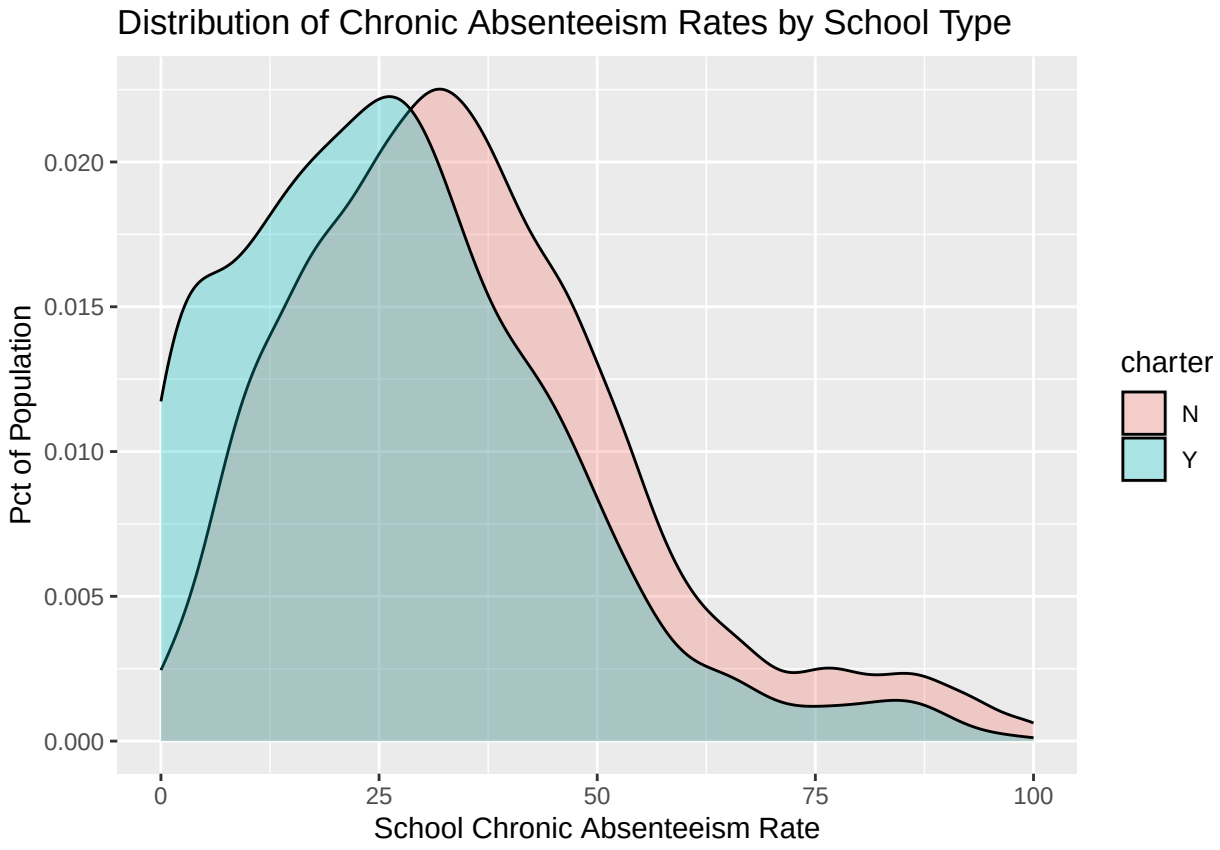
Finally, for chronic absenteeism rate we will use a t-test to determine if the difference in the average school chronic absenteeism rate between charter schools and public schools is statistically significant at the 95% level of significance

```
m5 <- merged_df %>%
  select(charter, status_type, school_level, chronic_absenteeism_rate)

m5 %>%
  count(charter, school_level) %>%
  spread(key=charter, value = n) %>%
  kable() %>%
  kable_material(c("striped"))
```

```
m5 <- m5 %>%
  filter(school_level != 'adult',
         school_level != 'preschool',
         school_level != 'ungraded')

m5 <- m5 %>%
  filter(!is.na(chronic_absenteeism_rate),
         str_detect(chronic_absenteeism_rate, regex('[[:alnum:]]'))) %>%
  mutate(chronic_absenteeism_rate = as.double(chronic_absenteeism_rate))

m5_summary <- m5 %>%
  group_by(charter) %>%
  summarize(n = n(),
            avg = round(mean(chronic_absenteeism_rate),2),
            sd = round(sd(chronic_absenteeism_rate),2))

ggplot(merged_df,aes(x=chronic_absenteeism_rate, fill=charter)) +
  geom_density(alpha=0.3) +
```

```
  ggtitle("Distribution of Chronic Absenteeism Rates by School Type") +
  xlab("School Chronic Absenteeism Rate") +
  ylab("Pct of Population")
```

```
## Warning: Removed 3955 rows containing non-finite values (‘stat_density()‘).
```

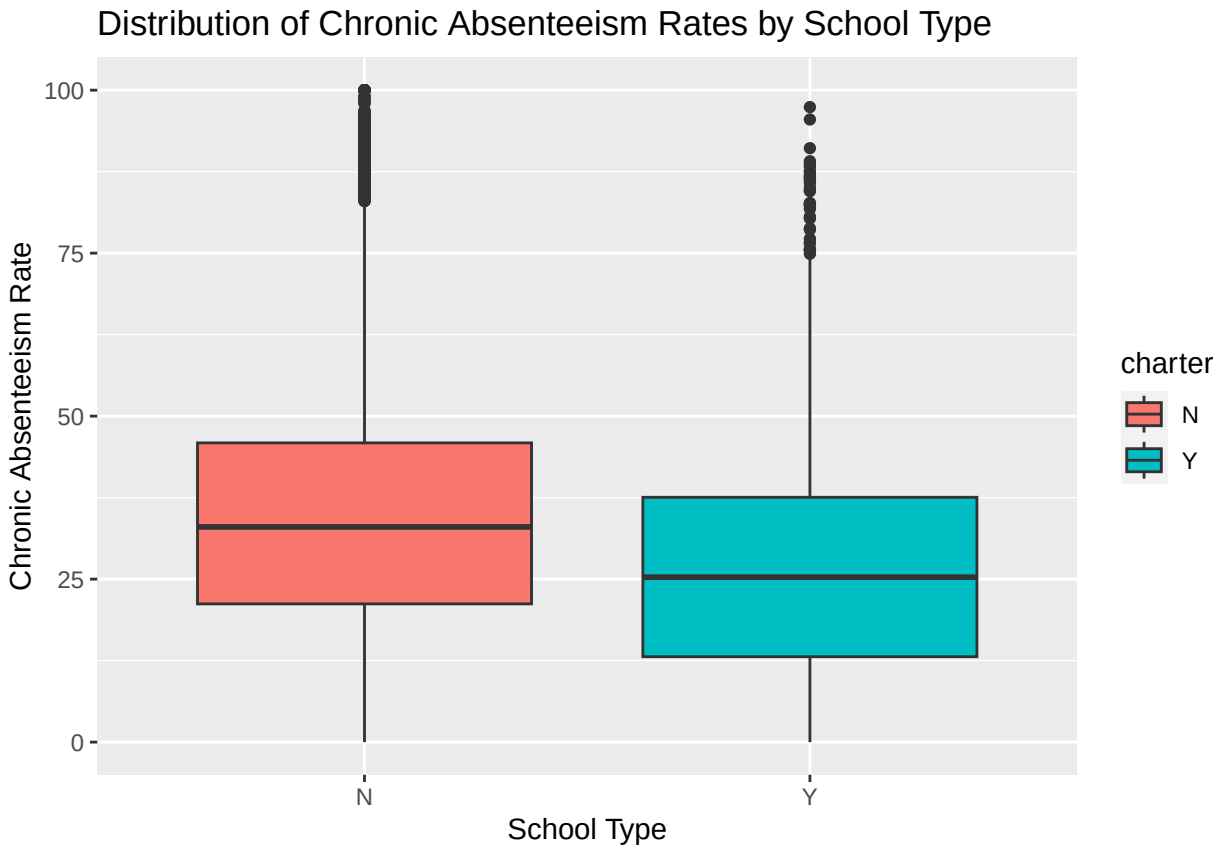



```
ggplot(merged_df) +
  geom_boxplot(aes(x=charter, y=chronic_absenteeism_rate, fill=charter)) +
  ggtitle("Distribution of Chronic Absenteeism Rates by School Type") +
  xlab("School Type") +
  ylab("Chronic Absenteeism Rate")
```

```
## Warning: Removed 3955 rows containing non-finite values (‘stat_boxplot()‘).
```

Table 10: Chronic Absenteeism Rate by School Type

| Charter School | N | Sample Mean | Sample Std. Dev |
|---|---|---|---|
| N | 8622 | 35.15 | 19.37 |
| Y | 1283 | 26.95 | 18.41 |




```
m5_summary %>%
  kable(
    caption = "Chronic Absenteeism Rate by School Type",
    col.names = c("Charter School", "N", "Sample Mean", "Sample Std. Dev")
  ) %>%
  kable_material(c("striped"))
```

```
ca_charter <- (m5 %>% filter(charter == 'Y'))$chronic_absenteeism_rate
ca_public <- (m5 %>% filter(charter == 'N'))$chronic_absenteeism_rate

t.test(x=ca_charter, y=ca_public, alternative="two.sided",conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  ca_charter and ca_public
## t = -14.78, df = 1732.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
```

Table 11: Charter School vs. Public Schools

| Category | Charter | Public |
|---|---|---|
| suspension_rate | 1 | 0 |
| graduation_rate | 0 | 1 |
| dropout_rate | 0 | 1 |
| college_going_rate | 0 | 0 |
| chronic_absenteeism_rate | 1 | 0 |

```
## 95 percent confidence interval:
##  -9.287698 -7.111556
## sample estimates:
## mean of x mean of y
##  26.94981  35.14943
```

```
chronic_absenteeism_rate = c(1,0)
```

At the 95% level of significance, we have enough evidence to **reject the null hypothesis** that there is no difference in the average chronic absenteeism rates between charter schools and pubic schools, and therefore we accept the alternative hypothesis that there is a difference in the average chronic absenteeism rates between charter schools and public schools.

Furthermore, based on our data, we conclude that the chronic absenteeism rate for charter schools is lower, and therefore better than the chronic absenteeism rate at public schools. Thus, we conclude that **charter schools perform better than public schools with respect to chronic absenteeism.**

# 6. Summary Data and Conclusion

Across each of the categories that we used to compare the performance of charter schools vs. public schools, we find that charter schools outperformed public schools in 2 of the 5 categories, while public schools outperformed charter schools in 2 of 5 categories.

```
category <- c("suspension_rate", "graduation_rate", "dropout_rate", "college_going_rate",
              "chronic_absenteeism_rate")
charter <- c(1,0,0,0,1)
public <- c(0,1,1,0,0)

summary_tab <- data.frame(category, charter, public)



summary_tab %>%
  kable(
    caption = 'Charter School vs. Public Schools',
    col.names = (c("Category", "Charter", "Public"))
  ) %>%
  kable_material(c("striped"))
```

Based on these results, we were unable to support our assumption that charter schools are better than public schools in California.

However, there are several known limitations with using this analysis to determine if charter schools are better than public schools. Some of these limitations include:

1. Based on the way we summarized the results from each of the different categories of analysis, we essentially are concluding that each category is equally-weighted for the purposes of calculating our overall ranking. While it's possible that an individual interested in comparing the two school types may agree that each of these categories are qualitatively the same in terms of their choice, additional research should be done to come to this conclusion.
2. Additionally, there are a number of other features that can and should be used to measure the differences between public and charter schools. Amongst them include: performance on standardized tests, extra-curriculur activities, diversity of staff, student diversity. Although one could argue that the performance in each of these other areas would show up in areas such as graduation rates and college-going rates, it's still worth noting that they are absent in our evaluation tool.
3. Finally, this analysis was based on relative performance between school types for all student. However, this doesn't take into account the fact that one of the missions of charter schools is often focused on beting a better option for undermarginalized populations. As a result, it's possible that if we were to further evaluate performance amongst different sub-cohorts of student populations, we may generate a different conclusion.

Overall, I believe that this research project allows us to recognize that there are no clear cut answers to the question of whether or not charter schools are better than public schools.

# Appendix:

The documentation for each of the datasets are located here

1. Public Schools and Districts - https://www.cde.ca.gov/ds/si/ds/fspubschls.asp
2. Chronic Absenteeism - https://www.cde.ca.gov/ds/ad/fsabd.asp
3. Suspensions - https://www.cde.ca.gov/ds/ad/fssd.asp
4. Dropout - https://www.cde.ca.gov/ds/ad/fsacgr.asp
5. Post-Secondary Enrollment - https://www.cde.ca.gov/ds/ad/fscgr16.asp