Michael Chen Project 2 Individual Report

Michael Chen

Introduction

This data set is the measurement of SAT performance for seniors in the schools of the state of Connecticut for 2012 and 2013. It recorded each school's district number and name, school name, number of test takers for each year, number of participants, number that met the benchmark, and their respective percentage change from 2012 to 2013.

I'll explore this data by comparing which schools performed the best on the SAT. I'll measure this by checking their percentage of their test takers that met the benchmark for both years. Then I'll try to see if there's any pattern between these schools that might differentiate them from other schools. I'll also compare how well these schools did compared to other schools. I'll also take a look at the lowest-scoring schools and see what patterns they exhibited that were different from other schools.

Next I'll explore the average percentage meeting benchmark for both years to see if there's any change. I'll observe each district separately, and also collectively. For the districts that had a drastic change in their average meeting benchmark, I'll try to find any patterns that might distinguish these districts from other districts.

Lastly, I'll try to estimate the total population of high school students in the state of Connecticut. I'll do this by estimating the total population of senior students for each school by multiplying the number of test takers by the the inverse of the participation rate, then multiply it by 4 to estimate the entire student body for each school. Then I can sum them all up to get the total population for the entire state of Connecticut. I'll also take a look to see if the population changed from 2012 to 2013.

Questions and findings:

Which schools had the most percent meeting benchmark in 2012 and 2013?

First we'll need to import the data

```r
data <- read_csv("SAT_School_Participation_and_Performance__2012-2013.csv")
```

We'll try to see which schools had the highest percentage meeting benchmark in 2012

```r
data %>%
  arrange(desc(`Percent Meeting Benchmark: 2012`)) %>%
  select(`Percent Meeting Benchmark: 2012`,School,District) %>%
  head()
```

```
## # A tibble: 6 x 3
##   `Percent Meeting Benchmark: 2012` School                 District
##                               <dbl> <chr>                  <chr>
## 1                                82 New Canaan High School New Canaan
## 2                                82 Wilton High School     Wilton
## 3                                81 Weston High School     Weston
## 4                                81 Staples High School    Westport
## 5                                80 Darien High School     Darien
## 6                                75 Canton High School     Canton
```

Table 1.1 Highest Percentage Meeting Benchmarks in 2012

The schools with the highest scoring benchmarks are New Canaan High School, Wilton High School, Weston High School, Staples Highschool, and Darien High School. The other schools have a meeting benchmark below 80. One thing to note is that each of these schools came from a different district.

How much better are these schools compared with the rest of the state of Connecticut? We can find out using a boxplot, and highlighting where the top 5 schools are placed.

```
#Create a new datatable for the top 5
top5 <- data %>%
  arrange(desc(`Percent Meeting Benchmark: 2012`)) %>%
  select(`Percent Meeting Benchmark: 2012`,School,District) %>%
  head(5)
ggplot(data) +
  geom_boxplot(aes("Percentage meeting Benchmark",`Percent Meeting Benchmark: 2012`)) +
  #overlap our plot with data from the top 5
  geom_point(data=top5,aes("Percentage meeting Benchmark",`Percent Meeting Benchmark: 2012`),color="Red
  theme_bw() +
  labs(title="Percentage meeting Benchmark in 2012",x="",y="Percentage",caption="Figure 1.2 Boxplot of 
```

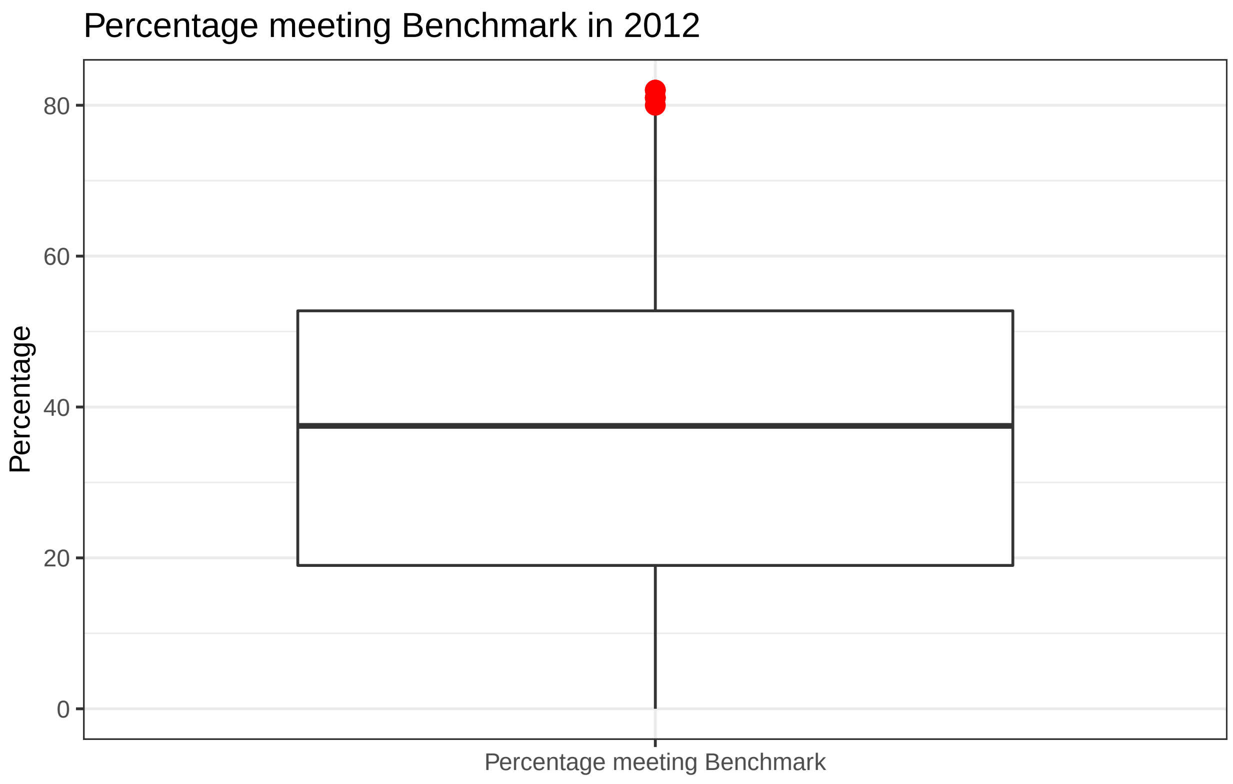

Figure 1.2 Boxplot of Percentage Meeting Benchmark in 2012, Top−5 Highlihted in Red

These schools scored significantly higher than all the other schools, they're right around the upper Inter-Quartile Range mark. However, a new question comes up while looking at this boxplot, it appears that some schools didn't have anyone meeting the benchmark. Let's see how many there are by looking at the distribution of schools for each percent meeting benchmark.

```
ggplot(data) +
  geom_histogram(aes(`Percent Meeting Benchmark: 2012`),na.rm=TRUE) +
  theme_bw() +
  scale_y_continuous(breaks=seq(0,15)) +
  labs(title="Count of schools for each percent meeting benchmark",caption="Figure 1.3")
```

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

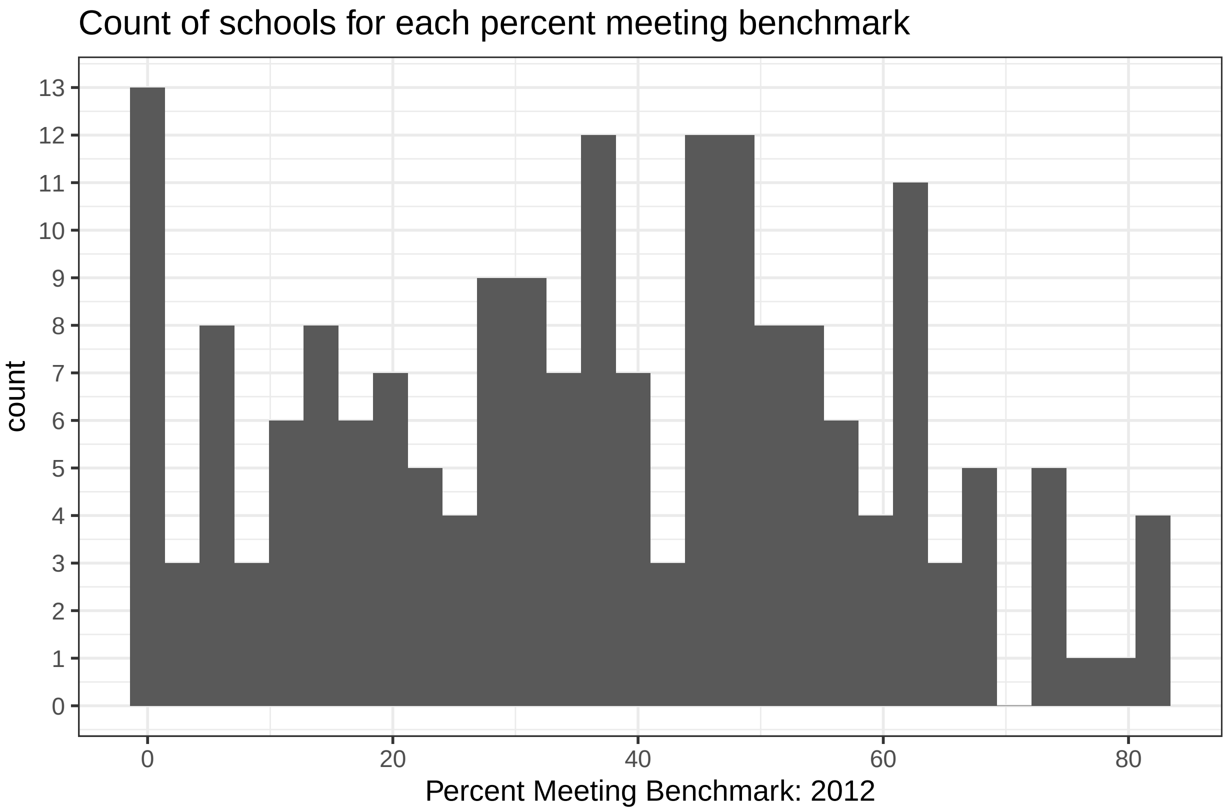

Figure 1.3

Around 13 schools had no one meeting the benchmark, let's find more information about them to see if we can find out why.

```
data %>%
  dplyr::filter(`Percent Meeting Benchmark: 2012`==0) %>%
  select(School,District,`Test-takers: 2012`,`Participation Rate (estimate): 2012`)
```

```
## # A tibble: 10 x 4
##    School              District      `Test-takers: 2~ `Participation Rate~
##    <chr>               <chr>                    <dbl>                <dbl>
##  1 Culinary Arts Acad~ Hartford                    30                   71
##  2 Hartford Public Hi~ Hartford                    82                   76
##  3 High School Inc.    Hartford                    28                   76
##  4 OPPortunity High S~ Hartford                    38                  100
##  5 Hyde Leadership Sc~ New Haven                   37                   86
##  6 Riverside Educatio~ New Haven                   12                  100
##  7 CREC - Public Safe~ Capitol Regio~               13                      87
##  8 Stamford Academy    Stamford Acad~               11                      28
##  9 Bullard Havens Tec~ Connecticut T~              127                      64
## 10 Eli Whitney High S~ Connecticut T~               76                      65
```

Table 1.4

The schools with no one meeting the benchmark happened particuarly in the districts of Hartford, New Haven, and Connecticut Technical High School System. There's nothing unusual about the number of test-takers or participation rate. Perhaps the schools themselves are just not doing well.

Now we'll redo this for 2013 and compare the results with 2012.

```
data %>%
  #We want to see the schools from the highest percentage meeting benchmark
  arrange(desc(`Percent Meeting Benchmark: 2013`)) %>%
  select(`Percent Meeting Benchmark: 2013`,School,District) %>%
  head()
```

```
## # A tibble: 6 x 3
##   `Percent Meeting Benchmark: 2013` School                 District
##                               <dbl> <chr>                  <chr>
## 1                                86 Darien High School     Darien
## 2                                84 Staples High School    Westport
## 3                                83 New Canaan High School New Canaan
## 4                                83 Weston High School     Weston
## 5                                81 Wilton High School     Wilton
## 6                                78 Ridgefield High School Ridgefield
```

Table 1.5 Highest Percentage Meeting Benchmarks in 2013

The schools with the highest percentage of benchmark being met are Darien High School, Staples High School, New Canaan High School, Weston High School, and Wilton High School. The other schools scored below 80 are not worth mentioning. Interestingly, these schools are the same top scorers as the ones in 2012.

Let's compare these schools with the other schools:

```
#Create a new dataframe for the top 5 schools
top5 <- data %>%
  arrange(desc(`Percent Meeting Benchmark: 2013`)) %>%
  select(`Percent Meeting Benchmark: 2013`,School,District) %>%
  head(5)
ggplot(data) +
  geom_boxplot(aes("Percentage meeting Benchmark",`Percent Meeting Benchmark: 2013`)) +
  #Overlap our  boxplot with the top5 schools for comparison
  geom_point(data=top5,aes("Percentage meeting Benchmark",`Percent Meeting Benchmark: 2013`),color="Red
  theme_bw() +
  labs(title="Percentage meeting Benchmark in 2013",x="",y="Percentage",caption="Figure 1.6 Boxplot of 
```

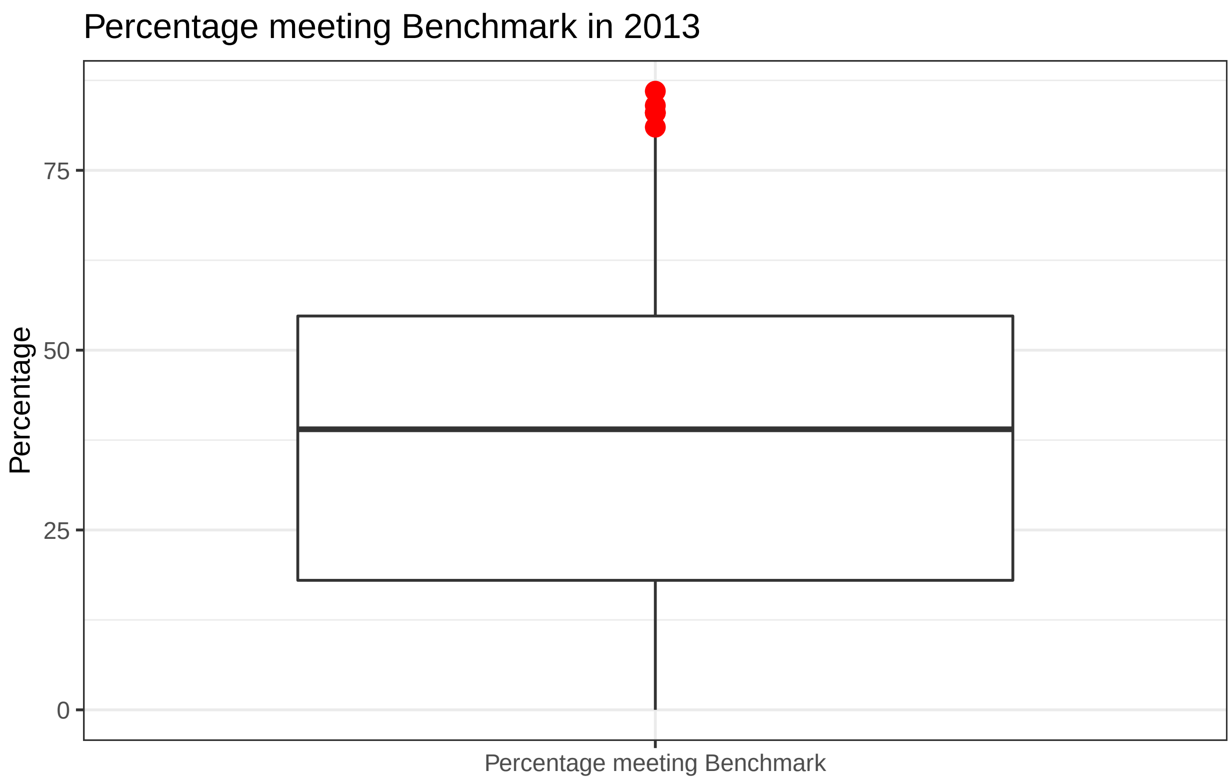

Figure 1.6 Boxplot of Percentage Meeting Benchmark in 2013 with highest−scorers highlighted in red

It appears we have the same situation as 2012, where the top 5 scorers are just around the upper inter-quartile range.

There are still some schools that had no one meeting the benchmark. Perhaps they are the same schools from the same district.

```
data %>%
  #Filter by the schools that had no one meeting the benchmark to get the poorest performing schools
  dplyr::filter(`Percent Meeting Benchmark: 2013`==0) %>%
  select(School,District)
```

```
## # A tibble: 7 x 2
##   School                                      District
##   <chr>                                       <chr>
## 1 Culinary Arts Academy                       Hartford
## 2 HPHS Nursing Academy                        Hartford
## 3 Hyde Leadership School                      New Haven
## 4 Riverside Education Academy                 New Haven
## 5 CREC - Public Safety Academy Interdistric~ Capitol Region Education Coun~
## 6 Eli Whitney High School                     Connecticut Technical High Sc~
## 7 Albert I Prince Technical High School       Connecticut Technical High Sc~
```

Table 1.7

Some schools are different, but they all came from the same district of Hartford, New Haven, and Connecticut Technical High School System.

Is the average percentage of meeting the benchmark higher in 2013 than 2012?

We'll compute the average meeting benchmark of 2012 then 2013 and compare them.

```
data %>%
  #Remove NA values so they won't affect our measurement
  summarize("2012 Mean Percentage Benchmark Meet" = mean(`Percent Meeting Benchmark: 2012`,na.rm=TRUE),
            "2013 Mean Percentage Benchmark Meet" = mean(`Percent Meeting Benchmark: 2013`,na.rm=TRUE)
            )
```

```
## # A tibble: 1 x 2
##   `2012 Mean Percentage Benchmark Mee~ `2013 Mean Percentage Benchmark Mee~
##                                  <dbl>                                <dbl>
## 1                                 36.5                                 37.1
```

Table 2.1 Mean percentage of benchmark meet for each year

There is indeed a change from 2012 to 2013, but not a lot. The scores in 2013 are about 0.52% higher than 2012, which is very little. Even though there's not much change, there may be some schools that have changed dramatically, let's try to discover the biggest change in percent meeting benchmark.

```
data %>%
  arrange(desc(`Percent Meeting Benchmark: Change%`)) %>%
  select(`Percent Meeting Benchmark: Change%`,School) %>%
  head(3)
```

```
## # A tibble: 3 x 2
##   `Percent Meeting Benchmark: Change%` School
##                                  <dbl> <chr>
## 1                                   20 Parish Hill High School
## 2                                   16 Lewis S Mills High School
## 3                                   12 Tourtellotte Memorial High School
```

Table 2.2 Schools with the most drastic change in percentage meeting benchmark

There is only one school that had a significant improvement of 20% increase in percentage of benchmark meets, which shot from 32% to 52%, and that is Parish Hill High School.

We can also compare the which districts had the biggest improvements in percent meeting benchmarks.

```
data %>%
  #We wish to look at each District
  group_by(District) %>%
  #Districts may have multiple schools so we'll take the average
  summarize(Avg_Benchmark_Change = round(mean(`Percent Meeting Benchmark: Change%`))) %>%
  arrange(desc(Avg_Benchmark_Change)) %>%
  #Only focus on the schools that are significant so they don't crowd our plot
  dplyr::filter(Avg_Benchmark_Change > 10) %>%
  ggplot() +
  geom_bar(aes(District,Avg_Benchmark_Change),stat="identity") +
  theme_bw() +
  labs(title="Most Improved Districts",caption="Figure 2.3")
```

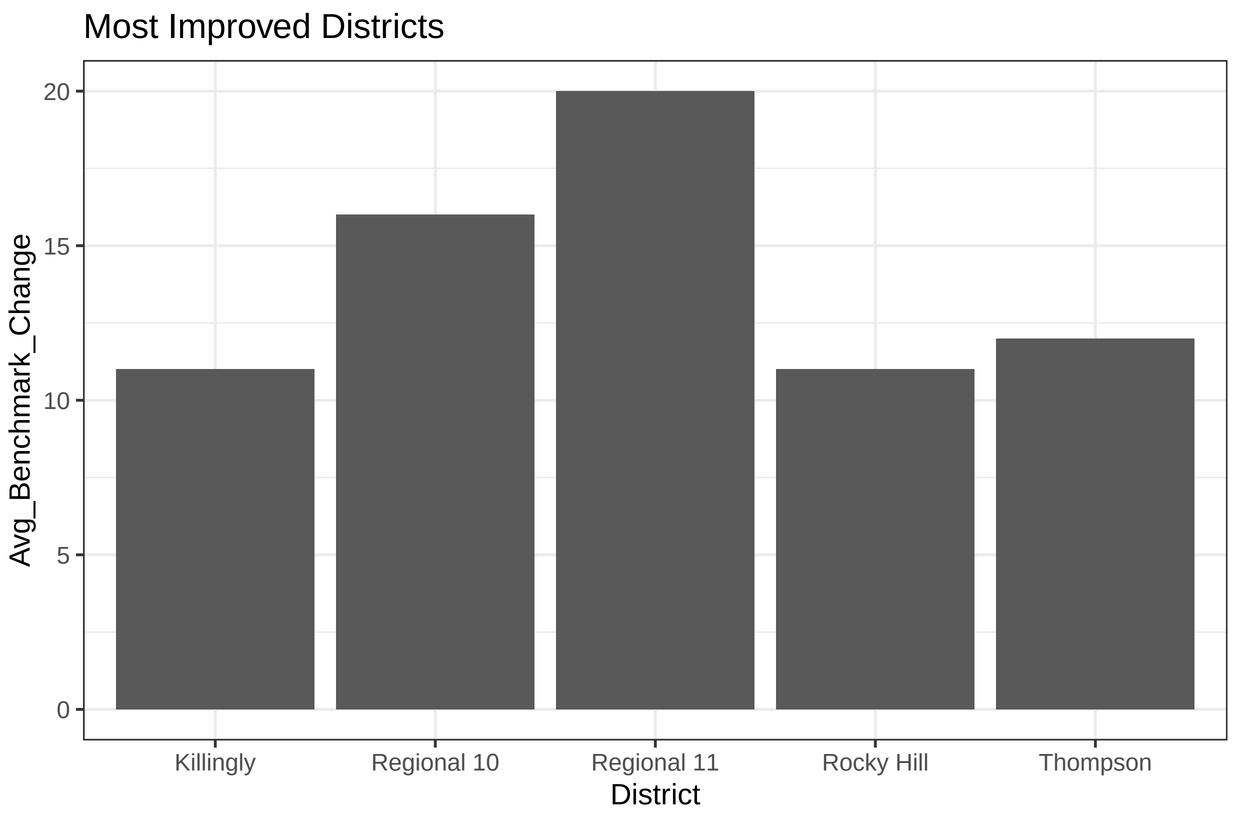

Figure 2.3

What is the population of high-school students in this state?

To answer this question, we'll need the number of test takers for each school, the participation rate of each school, and use that to find the student population of each school, and then we can sum that up to find the total population of students in Connecticut. One problem to note is that this data only provides the number of seniors of a highschool, and it find the total population for each highschool, I'm going to multiply the number of seniors by 4. This estimator might be a little biased since the number of students for the other grade-levels may not be the exact same as the number of senior students, but it is still a good estimator.

Let's first find an estimate for the population in 2012.

```
data %>%
  #4 times the number of test takers times the inverse of the participation rate
  transmute(student_population = 4*(`Test-takers: 2012`*(100 / `Participation Rate (estimate): 2012`)))
  summarize("2012 Total Student Population"=sum(student_population,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   `2012 Total Student Population`
##                             <dbl>
## 1                         158996.
```

Table 3.1 Estimate of the 2012 student population

There is around 160,000 high-school students in the state of Conneciticut in 2012. Now let's compare it with 2013.

```
data %>%
  transmute(student_population = 4*(`Test-takers: 2013`*(100 / `Participation Rate (estimate): 2013`)))
  summarize("2013 Student Population"=sum(student_population,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   `2013 Student Population`
##                       <dbl>
## 1                   160383.
```

Table 3.2 Estimate of the 2013 student population.

It is also around 160,000, and perhaps slightly higher, but there's virtually no change in the student population from 2012 to 2013.

Let's see which school has the most students.

```
data %>%
  #First add the estimated student population for each year
  mutate(student_population_2013 = 4*(`Test-takers: 2013`*(100 / `Participation Rate (estimate): 2013`))
  mutate(student_population_2012 = 4*(`Test-takers: 2012`*(100 / `Participation Rate (estimate): 2012`))
  arrange(desc(student_population_2012)) %>%
  select(School,District,student_population_2012,student_population_2013) %>%
  #Then round them
  mutate(student_population_2012 = round(student_population_2012,1),student_population_2013 = round(stu
  #We need to facet by year later on so we'll gather the columns of student_population2012 and 2013
  gather(key="Student Population Year",value="Student Population",student_population_2012,student_popula
  #Only look at the significant ones
  dplyr::filter(`Student Population` > 2000) %>%
  ggplot() +
  geom_bar(aes(School,`Student Population`,fill=School),stat="identity") +
  facet_wrap(~`Student Population Year`) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 50, hjust = 1)) +
  labs(title="Top Schools with the Most Number of Students",caption="Figure 3.3 schools with the highes
```

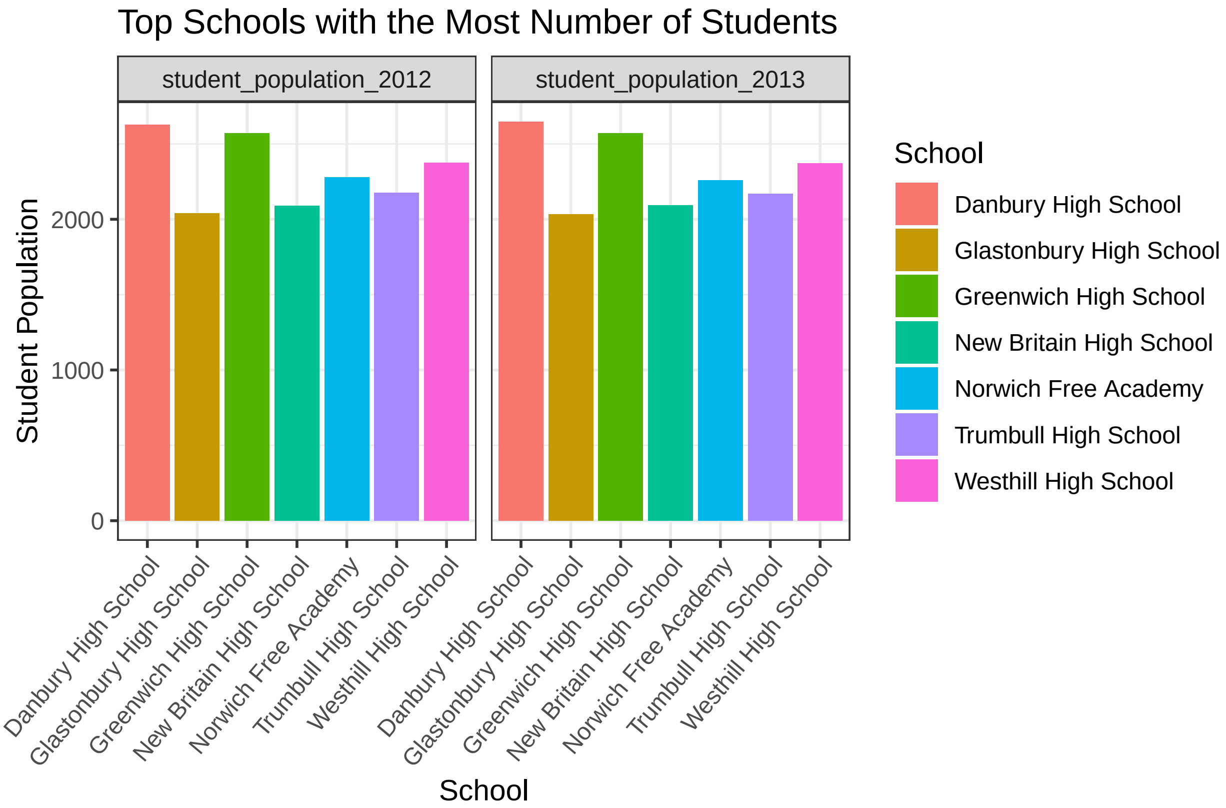

Figure 3.3 schools with the highest student population based on my estimator

The schools with the highest student population remained the same for both years, and they each have a population of over 2000.

Conclusion

Based on my findings, if you live in Connecticut, then it's best if you send your child to one of these top 5 highschools: New Canaan High School, Wilton High School, Weston High School, Staples Highschool, or Darien High School, since they've been on the top for both 2012 and 2013. If you can't live in one of thoes schools, you can also try to go to Parish Hill High School since it's fastest improving school, so there's also hope for your student to do well there in the future. If you can't make it to any of these schools, at least avoid the schools with the lowest-scoring benchmarks, namely Hartford, New Haven, and Connecticut Technical High School System.

It looks as if the average percentage meeting benchmark is about the same for both years, but it's very low. Some questions could be raised about why it's so low, or perhaps it's not very low compared to other states. In order to figure that out, we'd need more data from other states so we could compare Connecticut's average meeting benchmark with theirs. If we were to look at the individual districts, some regions have improved their percentage meeting benchmark greatly, like Regional 10 and 11, who both had an increase of at least 15%.

There's not much change in the student population of Connecitcut from 2012-2013, which hovered around 160,000. This number is just an estimate, and it could be far from the actual number for the population of high school students in Connecticut. One big disadvantage of my estimator is that I assumed the student body of seniors is equal to the student body of juniors, sophomores and freshman. It's likely that these groups of students don't always have the same number, but we don't know by how much, so we can mitigate this bias if we had data indicating the number of students for each of these groups.