achievement_first_technical_exercise.rmd

Drake Wagner

6/26/2020

```
#setwd('/home/drake/R')
setwd('C:\\Users\\dwagn\\Downloads\\R')
library("readxl")
library('dplyr')
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library('ggplot2')

school_data <- read_excel('F&P Sample Data Set.xlsx')
# ID, School name, Beginning of year score, End of year score

# Searches for variation within the values. I omitted the student IDs since
# every id would vary. This will allow us to easily see any misspellings or
# differences in how values were recorded (for example, "5.0" vs. "5th")
unique_values <- apply(school_data[, c('School Name',
                                       'Grade Level',
                                       'BOY F&P Score',
                                       'EOY F&P Score')], 2, unique)
# places the scores in numerical order and assigns them as integers
unique_values$`BOY F&P Score`=sort(as.integer(unique_values$`BOY F&P Score`))
unique_values$`EOY F&P Score`=sort(as.integer(unique_values$`EOY F&P Score`))
unique_values
```

```
## $`School Name`
## [1] "Bushwick Middle School"      "Crown Heights Middle School"
## [3] "Bushwick MS"                 "Crown Hghts Middle School"
##
## $`Grade Level`
## [1] "5.0" "6.0" "5th" "6th"
##
```

```
## $`BOY F&P Score`
##  [1]  0  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##
## $`EOY F&P Score`
##  [1]  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 33
```

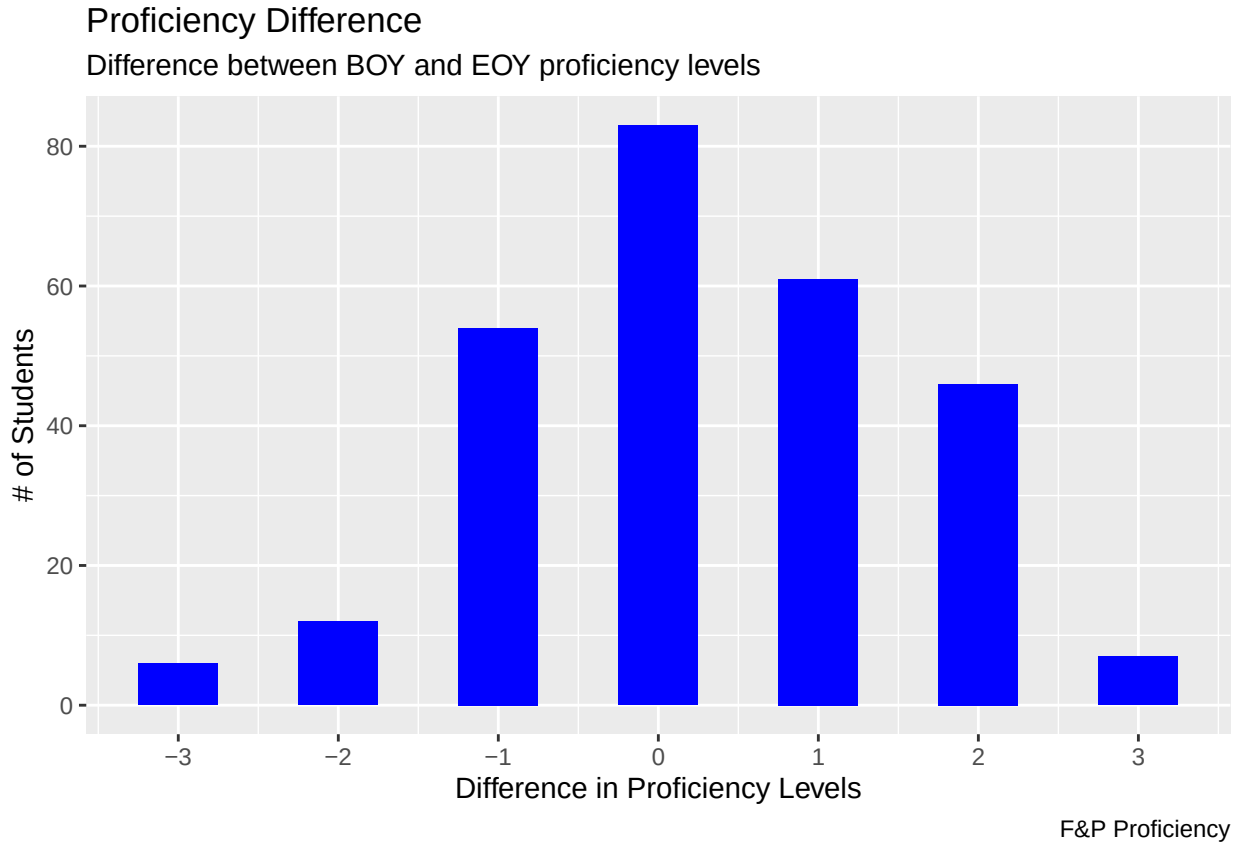
```
# renames the score columns to remove whitespace and special character from name
school_data <- school_data %>%
  rename(boy_score = 'BOY F&P Score',
         eoy_score = 'EOY F&P Score',
         school = 'School Name',
         grade = 'Grade Level')

# remove NA values
school_data <- na.omit(school_data)

# Since there are only a couple of values that need changing, I use grepl
# (similar to grep in shell scripts) to replace the values
school_data$grade[grepl(
  '5',school_data$grade)] <- '5.0'
school_data$grade[grepl(
  '6',school_data$grade)] <- '6.0'
school_data$school[grepl(
  'Bush',school_data$school)] <-'Bushwick Middle School'
school_data$school[grepl(
  'Crown',school_data$school)] <- 'Crown Heights Middle School'

# Another way this could have been done without additional packages:
#school_data$`Grade Level`[school_data$`Grade Level`=='5th'] <- '5.0'
#school_data$`Grade Level`[school_data$`Grade Level`=='6th'] <- '6.0'

# Now a quick double check to make sure the values have been replaced:
unique_values_2 <- apply(school_data[, c('school',
                                         'grade',
                                         'boy_score',
                                         'eoy_score')], 2, unique)
unique_values_2$`boy_score`=sort(as.integer(unique_values_2$`boy_score`))
unique_values_2$`eoy_score`=sort(as.integer(unique_values_2$`eoy_score`))
c(unique_values_2$school, unique_values_2$grade)
```

```
## [1] "Bushwick Middle School"      "Crown Heights Middle School"
## [3] "5.0"                         "6.0"
```

```
# potential 'ifelse' way of computing proficiency
# school_data <- school_data %>%
#   mutate(boy_proficiency = ifelse(boy_score<=9 & grade=='5.0', 'remedial',
#                            ifelse(boy_score<=11 & grade=='5.0', 'below proficient',
#                            ifelse(boy_score<= 13 & grade=='5.0', 'proficient',
#                            ifelse(boy_score>=14 & grade=='5.0', 'advanced',
#                            ifelse(boy_score<=11&grade=='6.0', 'remedial',
#                            ifelse(boy_score<=13&grade=='6.0', 'below proficient',
#                            ifelse(boy_score<=15&grade=='6.0', 'proficient',
#                            ifelse(boy_score>=16&grade=='6.0', 'advanced', ' '
```

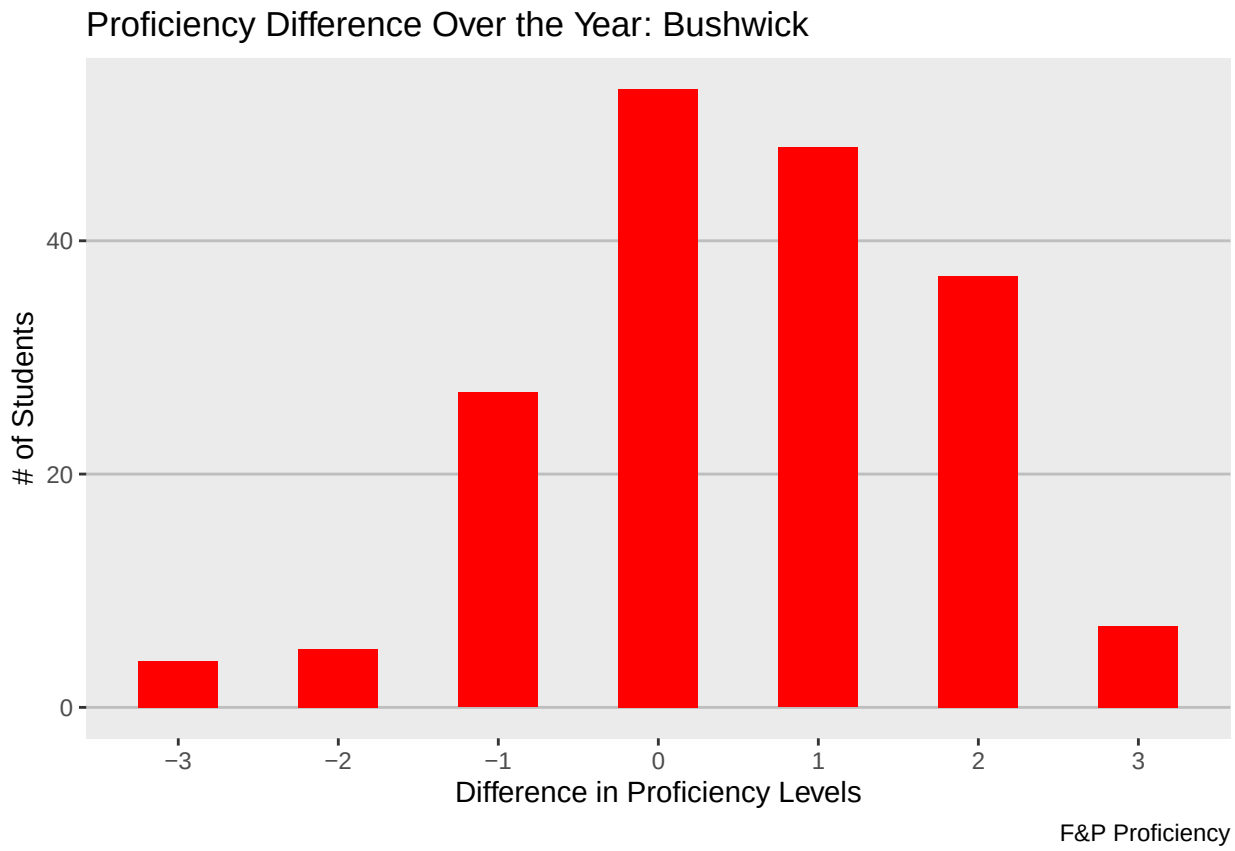
```r
# )))))))))

# Use dplyr to convert BOY and EOY scores to proficiency levels, according to
# the "F&P Proficiency Levels" tab
school_data <- school_data %>%
  mutate(boy_proficiency = case_when(
    boy_score <= 9 & grade == '5.0' ~ 'remedial',
    boy_score <= 11 & grade == '5.0' ~ 'below proficient',
    boy_score <= 13 & grade == '5.0' ~ 'proficient',
    boy_score >= 14 & grade == '5.0' ~ 'advanced',
    boy_score <= 11 & grade == '6.0' ~ 'remedial',
    boy_score <= 13 & grade == '6.0' ~ 'below proficient',
    boy_score <= 15 & grade == '6.0' ~ 'proficient',
    boy_score >= 16 & grade == '6.0' ~ 'advanced'
  ))

# Adds the is_prof column (1=proficient, 0=not proficient), according to eoy scores
school_data <- school_data %>%
  mutate(is_prof = case_when(
    eoy_score <15 & grade == '5.0' ~ 0,
    eoy_score >=15 & grade == '5.0'~ 1,
    eoy_score <18 & grade == '6.0' ~ 0,
    eoy_score >=18 & grade == '6.0' ~ 1
  ))


school_data <- school_data %>%
  mutate(eoy_proficiency = case_when(
    eoy_score <= 11 & grade == '5.0' ~ 'remedial',
    eoy_score <= 13 & grade == '5.0' ~ 'below proficient',
    eoy_score <= 15 & grade == '5.0' ~ 'proficient',
    eoy_score >= 16 & grade == '5.0' ~ 'advanced',
    eoy_score <= 13 & grade == '6.0' ~ 'remedial',
    eoy_score <= 15 & grade == '6.0' ~ 'below proficient',
    eoy_score <= 17 & grade == '6.0' ~ 'proficient',
    eoy_score >= 18 & grade == '6.0' ~ 'advanced'
  ))


# improvement column
# List the difference as an integer, where each integer is the difference
# of one proficiency level
school_data <- school_data %>%
  mutate(difference = case_when(
    boy_proficiency == eoy_proficiency ~ '0',
    boy_proficiency == 'remedial' & eoy_proficiency == 'below proficient' ~ '1',
    boy_proficiency == 'remedial' & eoy_proficiency == 'proficient' ~ '2',
    boy_proficiency == 'remedial' & eoy_proficiency == 'advanced' ~ '3',
    boy_proficiency == 'below proficient' & eoy_proficiency == 'remedial' ~ '-1',
    boy_proficiency == 'below proficient' & eoy_proficiency == 'proficient' ~ '1',
    boy_proficiency == 'below proficient' & eoy_proficiency == 'advanced' ~ '2',
    boy_proficiency == 'proficient' & eoy_proficiency == 'remedial' ~ '-2',
    boy_proficiency == 'proficient' & eoy_proficiency == 'below proficient' ~ '-1',
```

```
    boy_proficiency == 'proficient' & eoy_proficiency == 'advanced' ~ '1',
    boy_proficiency == 'advanced' & eoy_proficiency == 'remedial' ~ '-3',
    boy_proficiency == 'advanced' & eoy_proficiency == 'below proficient' ~ '-2',
    boy_proficiency == 'advanced' & eoy_proficiency == 'proficient' ~ '-1'
  ))
# set difference column as an integer for future calculations
school_data$difference = as.integer(school_data$difference)


# Differences
#
# plotting the differences in proficiency over the year
ggplot(school_data, aes(x=difference)) +
  geom_bar(width=.5, fill="blue") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  labs(title="Proficiency Difference",
       subtitle="Difference between BOY and EOY proficiency levels",
       y = '# of Students',
       x = 'Difference in Proficiency Levels',
       caption='F&P Proficiency')
```



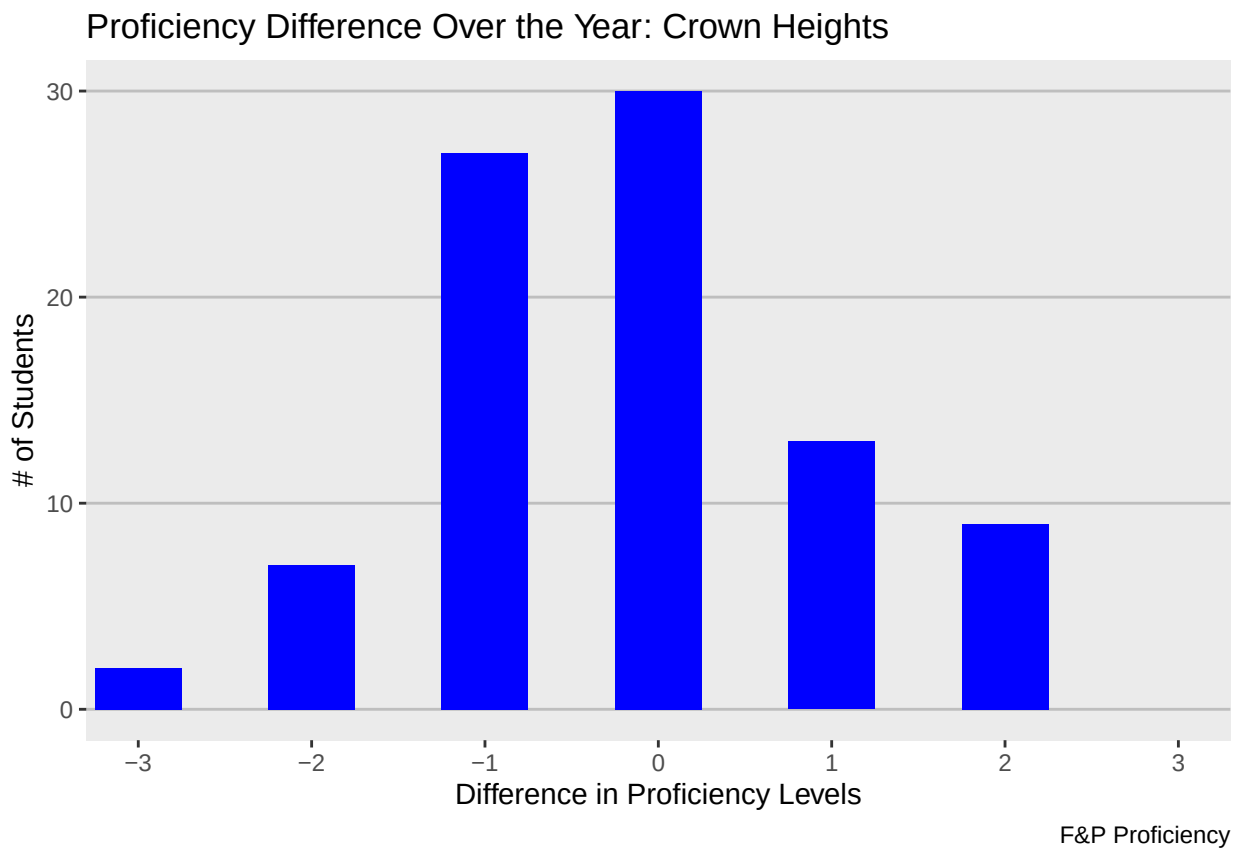

```
school_data_bushwick <- school_data %>%
  filter(school=="Bushwick Middle School")

school_data_crown <- school_data %>%
```

```
  filter(school=="Crown Heights Middle School")

# We can clearly see that this is skewed right
ggplot(school_data_bushwick, aes(x=difference)) +
  geom_bar(width=.5, fill="red") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 7)) +
  #xlim(-3, 3) +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.major.y = element_line(color='gray')) +
  labs(title="Proficiency Difference Over the Year: Bushwick",
       y = '# of Students',
       x = 'Difference in Proficiency Levels',
       caption='F&P Proficiency')
```

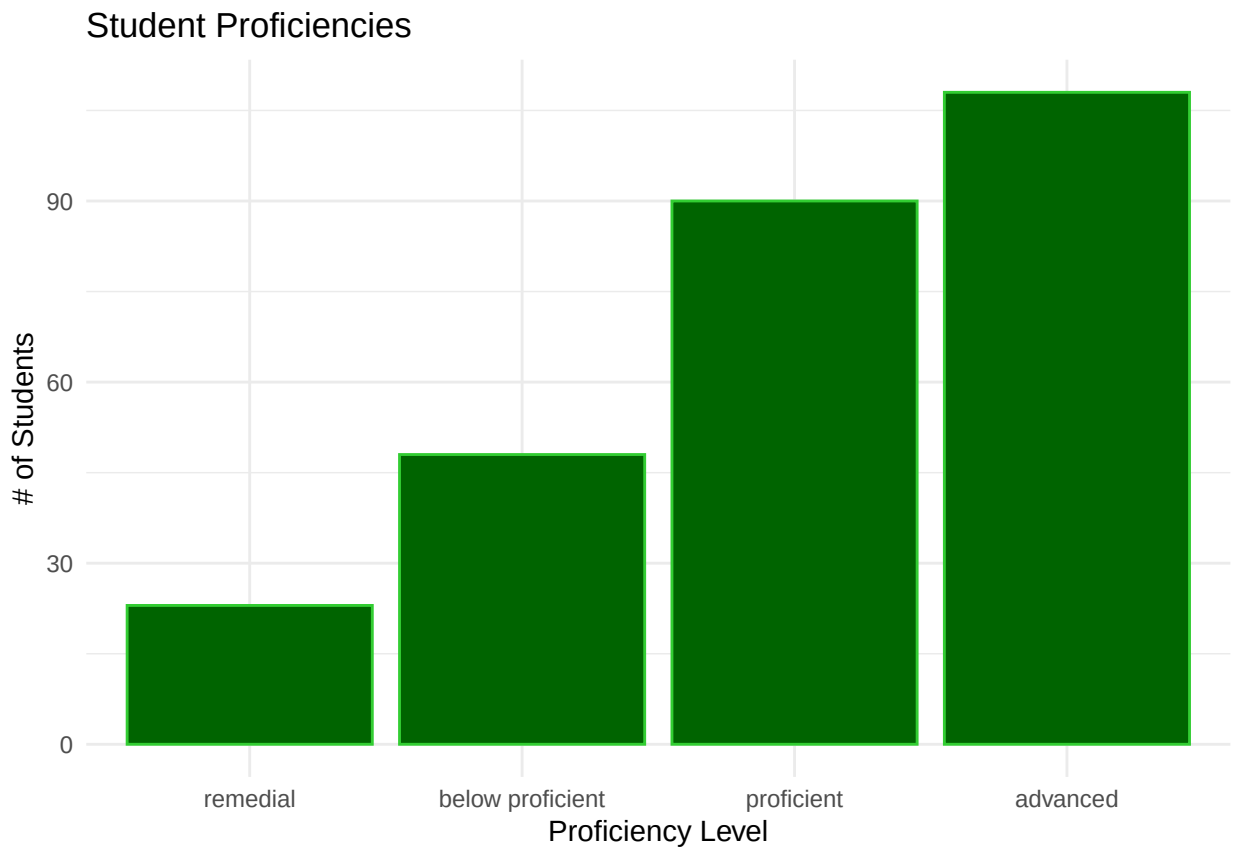



```
# Visually, this appears skewed slightly to the left
ggplot(school_data_crown, aes(x=difference)) +
  geom_bar(width=.5, fill="blue") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 7)) +
  coord_cartesian(xlim = c(-3, 3), clip = 'off') +
  #scale_x_discrete(limits = c(-3, -2, -1, 0, 1, 2, 3))
  #xlim() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.major.y = element_line(color='gray')) +
```

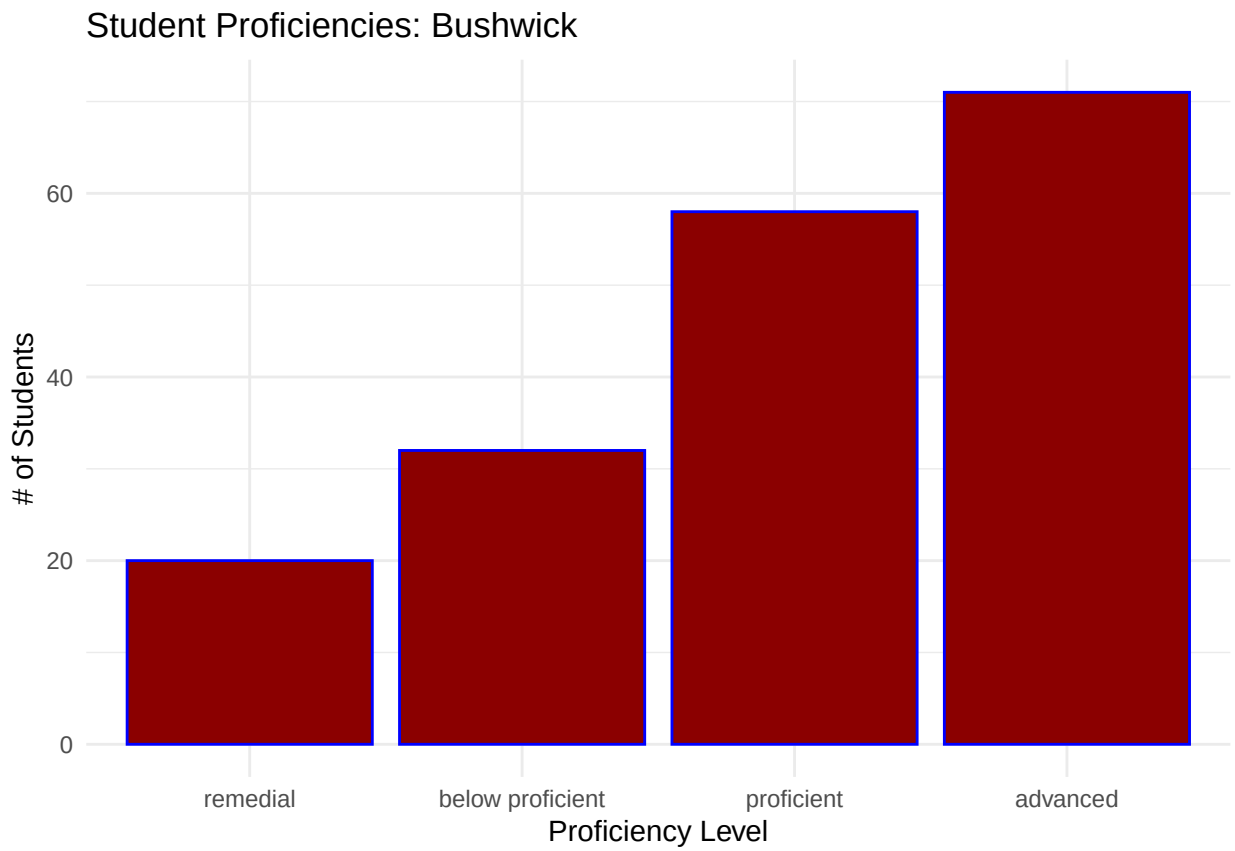
```
  labs(title="Proficiency Difference Over the Year: Crown Heights",
       y = '# of Students',
       x = 'Difference in Proficiency Levels',
       caption='F&P Proficiency')
```



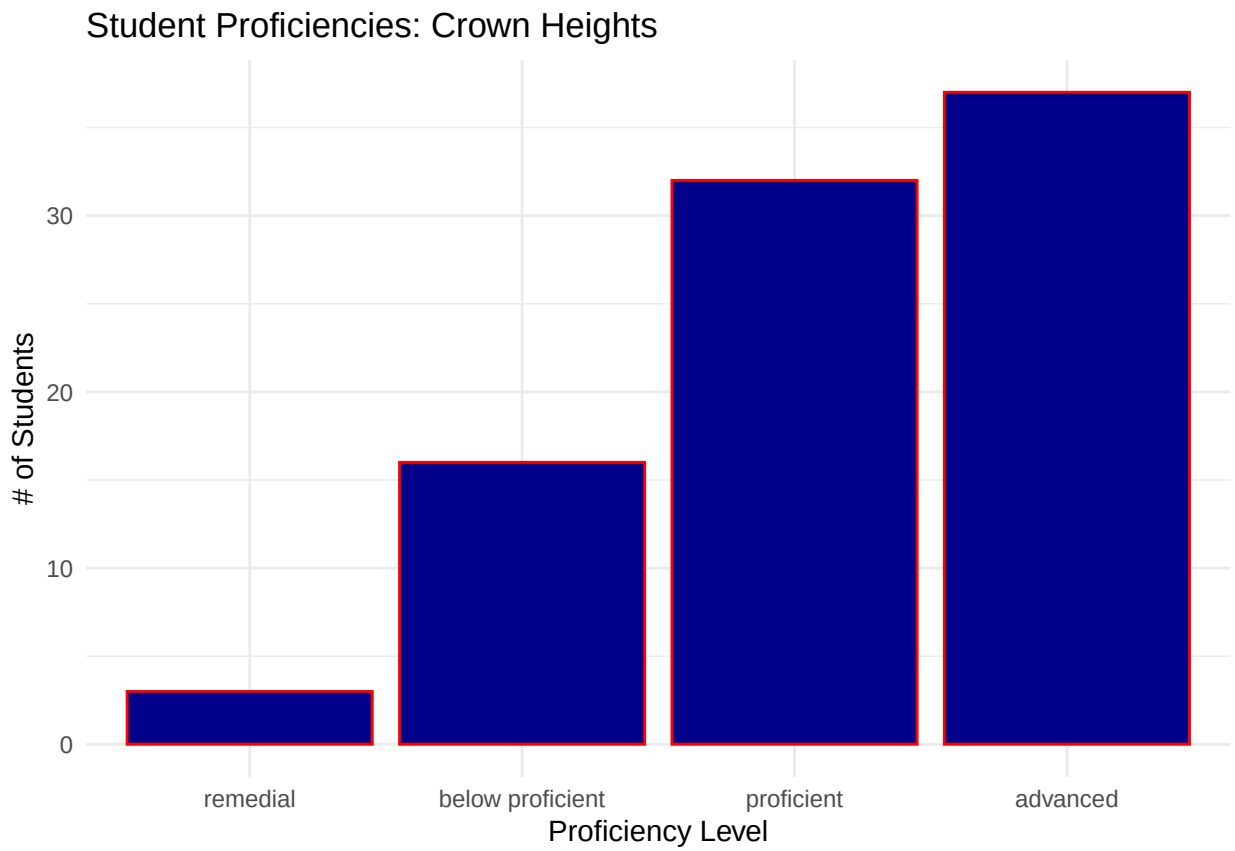

```
# Here I look more at final proficiency levels (eoy)

ggplot(school_data, aes(x=factor(school_data$eoy_proficiency, #Quick reordering of levels...
                                 levels = c('remedial', 'below proficient', 'proficient', 'advanced'))))
  labs(title='Student Proficiencies',
       x='Proficiency Level',
       y= '# of Students') +
  geom_bar(color="lime green", fill="dark green") +
  theme_minimal()
```




```
# And by school...

ggplot(school_data_bushwick, aes(x=factor(school_data_bushwick$eoy_proficiency, #Quick reordering of le
                                 levels = c('remedial', 'below proficient', 'proficient', 'advanced'))))
  labs(title='Student Proficiencies: Bushwick',
       x='Proficiency Level',
       y= '# of Students') +
  geom_bar(color='blue', fill="dark red") +
  theme_minimal()
```

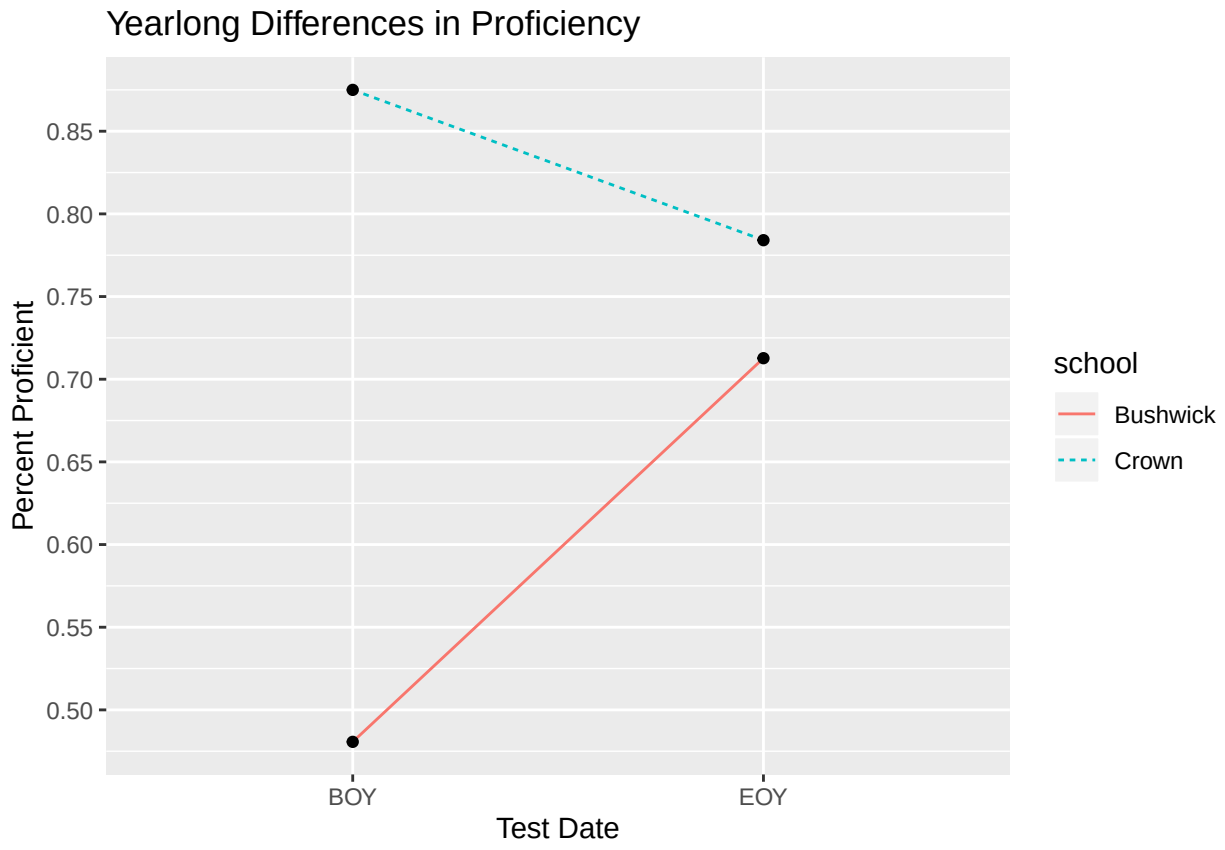



```
ggplot(school_data_crown, aes(x=factor(school_data_crown$eoy_proficiency, #Quick reordering of levels..
                                levels = c('remedial', 'below proficient', 'proficient', 'advanced'))))+
  labs(title='Student Proficiencies: Crown Heights',
       x='Proficiency Level',
       y= '# of Students') +
  geom_bar(color='red', fill="dark blue") +
  theme_minimal()
```



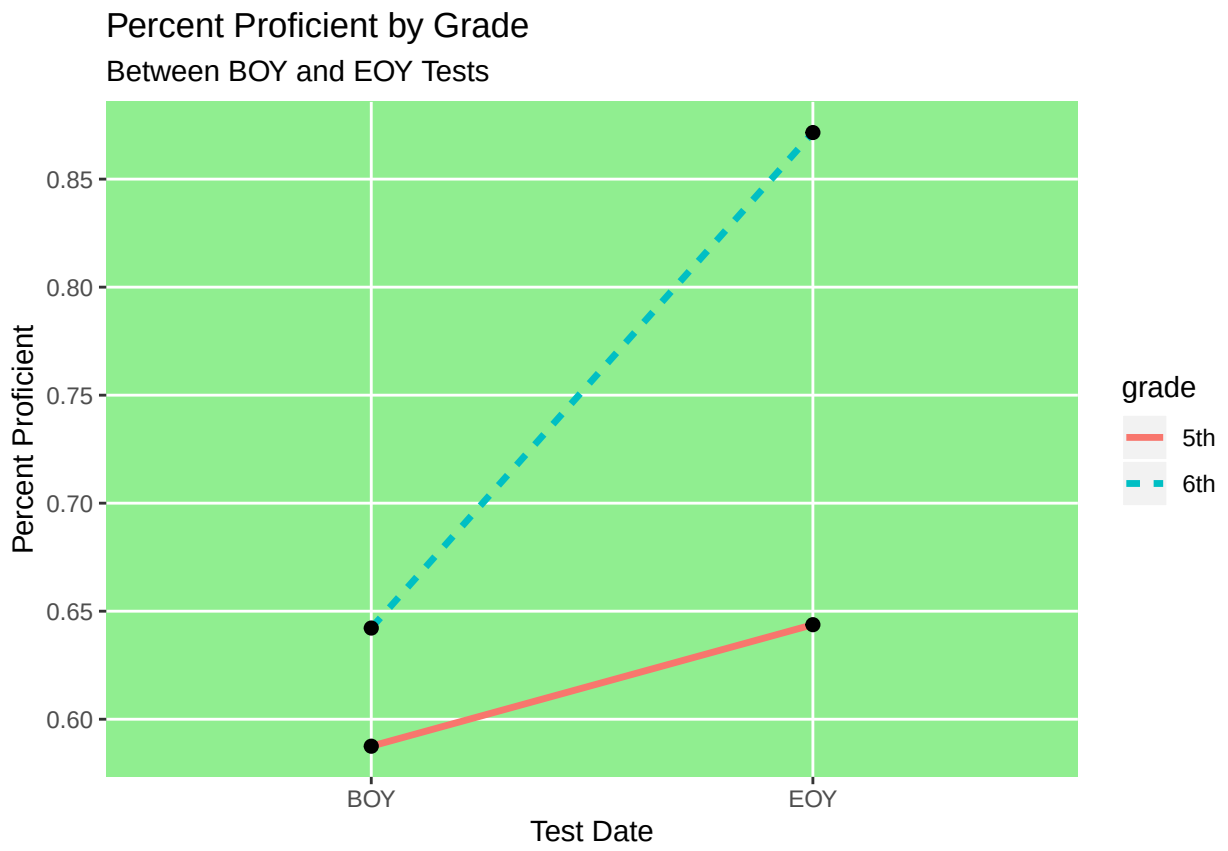

```
# Here, I take the two data frames I created of the separate schools and select
# only the student ids that have "proficient" or "advanced" scores recorded for
# the end of year tests, assuming EOY scores are our most current data...
num_prof_bush <- school_data_bushwick %>%
  filter(eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced')

num_prof_crown <- school_data_crown %>%
  filter(eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced')

# Now I calculate the percentage of students who recorded "proficient" or
# "advanced" end of year scores. We see that Crown has a higher percentage
# than Bushwick does (71.27% and 78.41%, respectively)
percent_prof_bush <- nrow(num_prof_bush)/nrow(school_data_bushwick)
percent_prof_crown <- nrow(num_prof_crown)/nrow(school_data_crown)

# same thing with boy
num_prof_bush_boy <- school_data_bushwick %>%
  filter(boy_proficiency == 'proficient' | boy_proficiency == 'advanced')

num_prof_crown_boy <- school_data_crown %>%
  filter(boy_proficiency == 'proficient' | boy_proficiency == 'advanced')

percent_prof_bush_boy <- nrow(num_prof_bush_boy)/nrow(school_data_bushwick)
percent_prof_crown_boy <- nrow(num_prof_crown_boy)/nrow(school_data_crown)

# Here I make a graph to visualize the improvement in test scores over the
```

```r
# course of the year, by school.
boy_eoy_by_school <- data.frame(school=rep(c('Bushwick', 'Crown'), each=2),
                                test=rep(c('BOY', 'EOY'), 2),
                                len=c(percent_prof_bush_boy, percent_prof_bush,
                                      percent_prof_crown_boy, percent_prof_crown))
ggplot(data=boy_eoy_by_school, aes(x=test, y=len, group=school)) +
  ylab('Percent Proficient') +
  xlab('Test Date') +
  labs(title='Yearlong Differences in Proficiency') +
  geom_line(aes(linetype=school, color=school)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  geom_point()
```



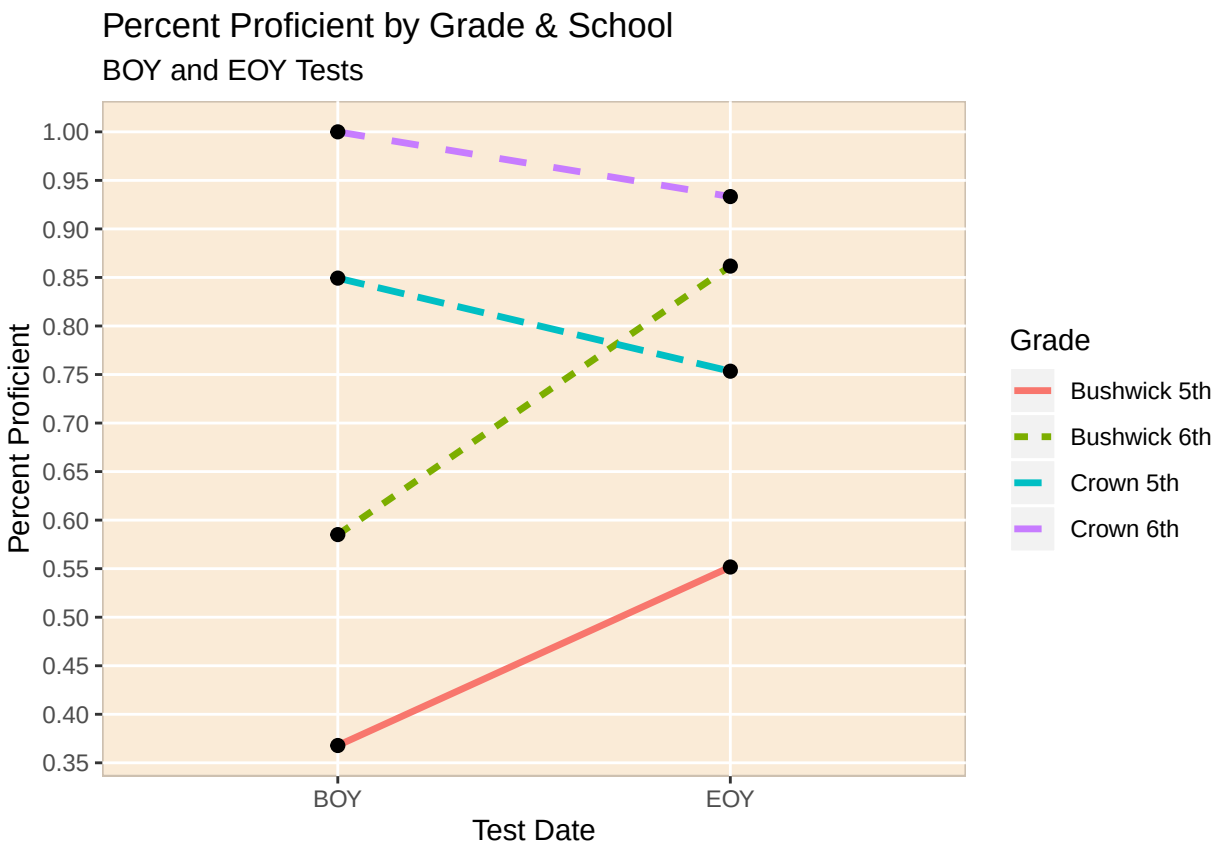

```r
# Now I want to see the trends according to the grade level. Again, I filter the
# dataset to only contain those who tested proficient or advanced, but instead of
# grouping by school, I group by grade
by_grade_5 <- school_data %>%
  filter(grade == '5.0')
by_grade_6 <- school_data %>%
  filter(grade == '6.0')


num_prof_5_boy <- school_data %>%
  filter((boy_proficiency == 'proficient' | boy_proficiency == 'advanced') & grade == '5.0')
num_prof_5_eoy <- school_data %>%
```

```r
  filter((eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced') & grade == '5.0')

# percent of 5th graders that are proficient, by test
percent_prof_5_boy <- nrow(num_prof_5_boy)/nrow(by_grade_5)
percent_prof_5_eoy <- nrow(num_prof_5_eoy)/nrow(by_grade_5)

num_prof_6_boy <- school_data %>%
  filter((boy_proficiency == 'proficient' | boy_proficiency == 'advanced') & grade == '6.0')
num_prof_6_eoy <- school_data %>%
  filter((eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced') & grade == '6.0')

# percent of 6th graders that are proficient, by test
percent_prof_6_boy <- nrow(num_prof_6_boy)/nrow(by_grade_6)
percent_prof_6_eoy <- nrow(num_prof_6_eoy)/nrow(by_grade_6)


# Now I combine the procifiency percent of 5th and 6th grade into a dataset, so that we can visualize i
boy_eoy_by_grade <- data.frame(grade=rep(c('5th', '6th'), each=2),
                               test=rep(c('BOY', 'EOY'), 2),
                               len=c(percent_prof_5_boy, percent_prof_5_eoy,
                                     percent_prof_6_boy, percent_prof_6_eoy))

ggplot(data=boy_eoy_by_grade, aes(x=test, y=len, group=grade)) +
  ylab('Percent Proficient') +
  xlab('Test Date') +
  labs(title='Percent Proficient by Grade',
       subtitle='Between BOY and EOY Tests')+
  geom_line(aes(linetype=grade, color=grade), size=1.2) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme(panel.background = element_rect(fill = "lightgreen",
                                        color = "lightgreen",
                                        size = 0.5, linetype = "solid"),
                                        panel.grid.minor = element_blank()) + # removes grid subsets
  geom_point(size=2)
```




```
# Separated by boy and eoy proficiency, school, and grade
# This is very messy, but is the data separated by all of those categories, testing the percent
# proficient at the time of each test
B5 <- school_data %>%
  filter(grade == '5.0' & school=='Bushwick Middle School')
B5boy <- nrow(B5 %>%
                filter(boy_proficiency == 'proficient' | boy_proficiency == 'advanced'))/nrow(B5) #perc
B5eoy <- nrow(B5 %>%
                filter(eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced'))/nrow(B5) #perc


B6 <- school_data %>%
  filter(grade == '6.0'& school=='Bushwick Middle School')
B6boy <- nrow(B6 %>%
                filter(boy_proficiency == 'proficient' | boy_proficiency == 'advanced'))/nrow(B6) #perc
B6eoy <- nrow(B6 %>%
                filter(eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced'))/nrow(B6) #perc


CH5 <- school_data %>%
  filter(grade == '5.0' & school=='Crown Heights Middle School')
CH5boy <- nrow(CH5 %>%
                 filter(boy_proficiency == 'proficient' | boy_proficiency == 'advanced'))/nrow(CH5)
CH5eoy <- nrow(CH5 %>%
                 filter(eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced'))/nrow(CH5)
```

```
CH6 <- school_data %>%
  filter(grade == '6.0'& school=='Crown Heights Middle School')
CH6boy <- nrow(CH6 %>%
                 filter(boy_proficiency == 'proficient' | boy_proficiency == 'advanced'))/nrow(CH6)
CH6eoy <- nrow(CH6 %>%
                 filter(eoy_proficiency == 'proficient' | eoy_proficiency == 'advanced'))/nrow(CH6)

school_and_grade <- data.frame(Grade=rep(c('Bushwick 5th','Crown 5th', 'Bushwick 6th', 'Crown 6th'), ea
                               test=rep(c('BOY', 'EOY'), 2),
                               pprof=c(B5boy, B5eoy, CH5boy, CH5eoy,
                                       B6boy, B6eoy, CH6boy, CH6eoy))

ggplot(data=school_and_grade, aes(x=test, y=pprof, group=Grade)) +
  geom_line(aes(linetype=Grade, color=Grade), size=1.2) +
  ylab('Percent Proficient') +
  xlab('Test Date') +
  labs(title='Percent Proficient by Grade & School',
       subtitle='BOY and EOY Tests')+
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme(panel.background = element_rect(fill = "antiquewhite",
                                        color = "antiquewhite3",
                                        size = 0.5, linetype = "solid"),
        panel.grid.minor = element_blank()) + # removes grid subsets
  geom_point(size=2)
```



```r
# Further analysis shows a low, positive correlation between the student's grade and their difference.
# 6th graders tend to score better than 5th graders
cor(as.integer(school_data$grade), school_data$difference)
```

```
## [1] 0.2446995
```

```r
# What if we look at each school individually...
cor(as.integer(school_data_bushwick$grade), school_data_bushwick$difference)
```

```
## [1] 0.2361747
```

```r
cor(as.integer(school_data_crown$grade), school_data_crown$difference)
```

```
## [1] -0.007084734
```

```r
median(school_data_bushwick$boy_score)
```

```
## [1] 13
```

```r
median(school_data_crown$boy_score)
```

```
## [1] 14
```

```r
median(school_data_bushwick$eoy_score)
```

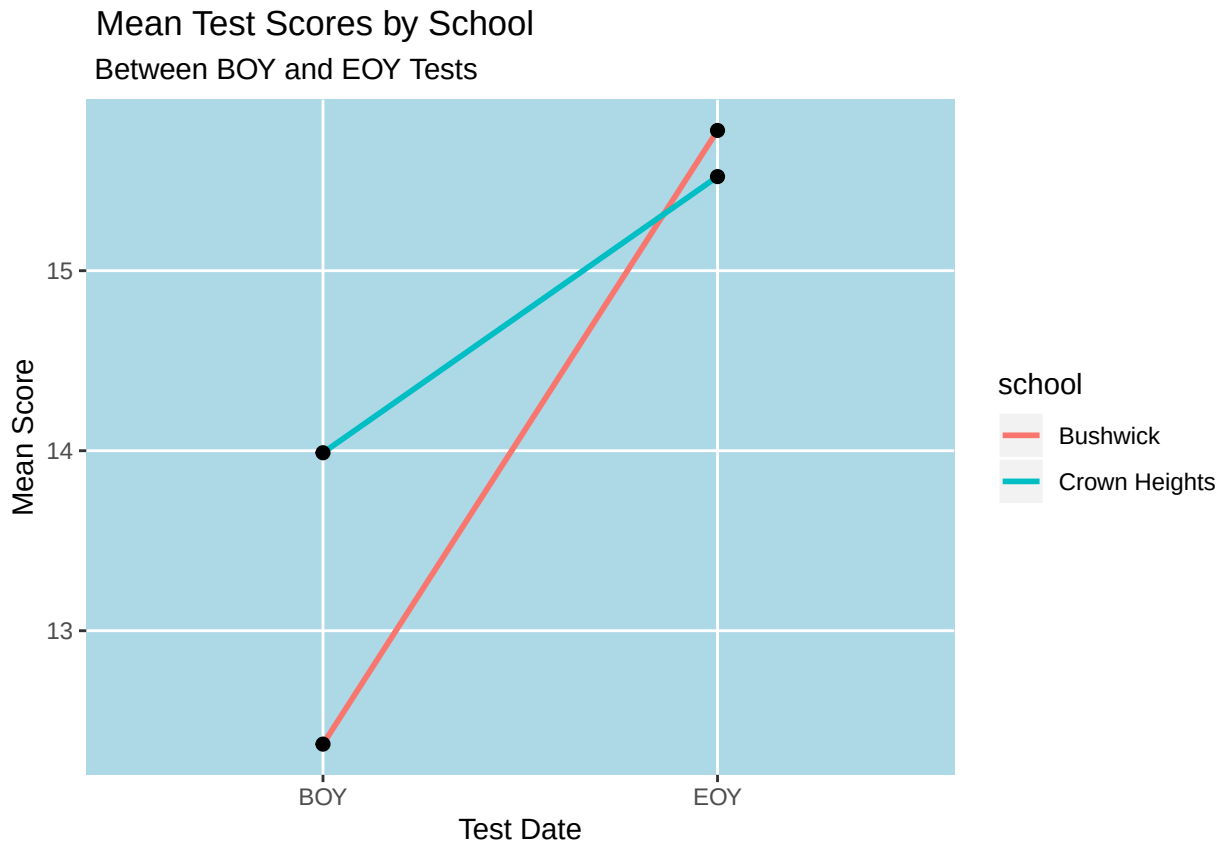
```
## [1] 16
```

```r
median(school_data_crown$eoy_score)
```

```
## [1] 15
```

```r
# While Bushwick's students test scores are positively correlated with their grade level, there appears
# no correlation within Crown's student body.
# Furthermore, despite Crown having a slightly higher median BOY test score than Bushwick, it appears t
# is greater than Crown's (median test score up 3 over the year at Bushwick vs. up 1 over the year at C
# once again assuming that the EOY test is taken with the same students at the end of their school year


# Here I make another graph with almost identical syntax to the last, comparing
# the mean scores by school this time rather than by grade. This will be a more
# helpful comparison in answering the proposed questions
meds <- data.frame(school=rep(c('Bushwick', 'Crown Heights'), each=2),
                           test=rep(c('BOY', 'EOY'), 2),
                                len=c(mean(school_data_bushwick$boy_score), mean(school_data_bushwick$eo
                                      mean(school_data_crown$boy_score), mean(school_data_crown$eoy_scor


ggplot(data=meds, aes(x=test, y=len, group=school)) +
```

```
  labs(title=" Mean Test Scores by School",
       subtitle=' Between BOY and EOY Tests',
       y = 'Mean Score',
       x = 'Test Date') +
  geom_line(aes(color=school), size=1) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 4)) +
  theme(panel.background = element_rect(fill = "lightblue",
                                        color = "lightblue",
                                        size = 0.5, linetype = "solid"),
                                        panel.grid.minor = element_blank()) +
  geom_point(size=2)
```




```
# Two-sample T-test
# Find is difference in EOY scores is significantly different between the two schools
# I finally use the "is_prof" column for this particular test. Once again, I assume that
# final proficiency is judged solely on the end of the year test

t.test(school_data_bushwick$is_prof, school_data_crown$is_prof, alternative='two.sided', var.equal=FALSE
```

```
##
##  Welch Two Sample t-test
##
## data:  school_data_bushwick$is_prof and school_data_crown$is_prof
## t = -2.3605, df = 178.05, p-value = 0.01933
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.27572442 -0.02462717
## sample estimates:
## mean of x mean of y
## 0.4861878 0.6363636
```

```
# While this is likely the most suitable type of significance testing that can be conducted between
# the two groups, it cannot be taken too seriously, since these are not random samples from a larger po
# However, assuming the samples from the two schools are independent without variance in the population
# there does appear to be a significant difference between the two schools in procifiency, p<.05.

nrow(school_data_bushwick)
```

```
## [1] 181
```

```
nrow(school_data_crown)
```

```
## [1] 88
```

```
percent_prof_bush
```

```
## [1] 0.7127072
```

```
percent_prof_crown
```

```
## [1] 0.7840909
```

```
head(school_data)
```

```
## # A tibble: 6 x 9
##   `Student ID` school grade boy_score eoy_score boy_proficiency is_prof
##          <dbl> <chr>  <chr>     <dbl>     <dbl> <chr>             <dbl>
## 1     10000001 Bushw~ 5.0          11        16 below proficie~       1
## 2     10000002 Bushw~ 5.0          11        16 below proficie~       1
## 3     10000003 Crown~ 5.0          11        16 below proficie~       1
## 4     10000004 Bushw~ 5.0          11        16 below proficie~       1
## 5     10000005 Bushw~ 5.0          11        14 below proficie~       0
## 6     10000006 Bushw~ 5.0          11        10 below proficie~       0
## # ... with 2 more variables: eoy_proficiency <chr>, difference <int>
```

```
#percent improvement within the schools
nrow(school_data_bushwick %>%
  filter(difference > 0)) # 92 improved (n=181)
```

```
## [1] 92
```

```
nrow(school_data_bushwick %>%
  filter(difference == 0)) # 53 no difference
```

```
## [1] 53
```

```r
nrow(school_data_bushwick %>%
  filter(difference < 0)) # 36 regressed
```

```
## [1] 36
```

```r
nrow(school_data_crown %>%
       filter(difference > 0)) # 22 improved (n=88)
```

```
## [1] 22
```

```r
nrow(school_data_crown %>%
       filter(difference == 0)) # 30 no difference
```

```
## [1] 30
```

```r
nrow(school_data_crown %>%
       filter(difference < 0)) # 36 regressed
```

```
## [1] 36
```