Liu_Milestone1

2023-11-07

Part I – Exploring

```
library(readxl)
library(tidyr)
library(ggplot2)
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(ggthemes)
```

loading and combining data

```
# Load dataset - FSA score from 2007/2008 - 2016/2017
FSA0716 <- read_excel("foundational_skills_assessment_2007-08_to_2016-17_residents_only.xlsx")
print(FSA0716)
```

```
## # A tibble: 18,777 x 15
##    SCHOOL_YEAR DATA_LEVEL    PUBLIC_OR_INDEPENDENT DISTRICT_NUMBER DISTRICT_NAME
##    <chr>       <chr>         <chr>                 <chr>           <chr>
##  1 2007/2008   Province Lev~ Province-Total        <NA>            <NA>
##  2 2007/2008   Province Lev~ Province-Total        <NA>            <NA>
##  3 2007/2008   Province Lev~ Province-Total        <NA>            <NA>
##  4 2007/2008   Province Lev~ Province-Total        <NA>            <NA>
##  5 2007/2008   Province Lev~ Province-Total        <NA>            <NA>
##  6 2007/2008   Province Lev~ BC Public School      <NA>            <NA>
##  7 2007/2008   Province Lev~ BC Public School      <NA>            <NA>
##  8 2007/2008   Province Lev~ BC Public School      <NA>            <NA>
##  9 2007/2008   Province Lev~ BC Public School      <NA>            <NA>
## 10 2007/2008   Province Lev~ BC Public School      <NA>            <NA>
## # i 18,767 more rows
## # i 10 more variables: SUB_POPULATION <chr>, GRADE <dbl>, FSA_SKILL_CODE <chr>,
## #   NUMBER_EXPECTED_WRITERS <chr>, NUMBER_WRITERS <chr>, NUMBER_UNKNOWN <chr>,
## #   NUMBER_EMERGING <chr>, NUMBER_ONTRACK <chr>, NUMBER_EXTENDING <chr>,
## #   SCORE <chr>
```

```r
# Load dataset - FSA score from 2017/2018 - 2020/2021
FSA1721 <- read_excel("foundational_skills_assessment_2017-18_to_2020-21_residents_only.xlsx")
print(FSA1721)
```

```
## # A tibble: 7,533 x 15
##    SCHOOL_YEAR DATA_LEVEL    PUBLIC_OR_INDEPENDENT DISTRICT_NUMBER DISTRICT_NAME
##    <chr>       <chr>         <chr>                 <chr>           <chr>
##  1 2017/2018   Province Lev~ Province-Total        <NA>            <NA>
##  2 2017/2018   Province Lev~ Province-Total        <NA>            <NA>
##  3 2017/2018   Province Lev~ Province-Total        <NA>            <NA>
##  4 2017/2018   Province Lev~ Province-Total        <NA>            <NA>
##  5 2017/2018   Province Lev~ Province-Total        <NA>            <NA>
##  6 2017/2018   Province Lev~ BC Public School      <NA>            <NA>
##  7 2017/2018   Province Lev~ BC Public School      <NA>            <NA>
##  8 2017/2018   Province Lev~ BC Public School      <NA>            <NA>
##  9 2017/2018   Province Lev~ BC Public School      <NA>            <NA>
## 10 2017/2018   Province Lev~ BC Public School      <NA>            <NA>
## # i 7,523 more rows
## # i 10 more variables: SUB_POPULATION <chr>, GRADE <dbl>, FSA_SKILL_CODE <chr>,
## #   NUMBER_EXPECTED_WRITERS <chr>, NUMBER_WRITERS <chr>, NUMBER_UNKNOWN <chr>,
## #   NUMBER_EMERGING <chr>, NUMBER_ONTRACK <chr>, NUMBER_EXTENDING <chr>,
## #   SCORE <chr>
```

```r
# Combine the two datasets "FSA0716" and "FSA1721" into one dataset called "FSA"
FSA <- rbind(FSA0716,FSA1721)
```

```r
# Compute descriptive statistics
summary(FSA)
```

```
##  SCHOOL_YEAR         DATA_LEVEL        PUBLIC_OR_INDEPENDENT DISTRICT_NUMBER
##  Length:26310       Length:26310       Length:26310          Length:26310
##  Class :character   Class :character   Class :character      Class :character
##  Mode  :character   Mode  :character   Mode  :character      Mode  :character
##
##
##
##  DISTRICT_NAME      SUB_POPULATION         GRADE       FSA_SKILL_CODE
##  Length:26310       Length:26310       Min.   :4.000   Length:26310
##  Class :character   Class :character   1st Qu.:4.000   Class :character
##  Mode  :character   Mode  :character   Median :7.000   Mode  :character
##                                        Mean   :5.501
##                                        3rd Qu.:7.000
##                                        Max.   :7.000
##  NUMBER_EXPECTED_WRITERS NUMBER_WRITERS     NUMBER_UNKNOWN
##  Length:26310            Length:26310       Length:26310
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##  NUMBER_EMERGING    NUMBER_ONTRACK     NUMBER_EXTENDING      SCORE
##  Length:26310       Length:26310       Length:26310       Length:26310
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
```

```
# See if there's any NA value in this dataset
sum(is.na(FSA))
```

```
## [1] 2448
```

```
# Display the names of columns with missing values
columns_with_na <- colnames(FSA)[colSums(is.na(FSA)) > 0]
print(columns_with_na)
```

```
## [1] "DISTRICT_NUMBER" "DISTRICT_NAME"
```

Cleaning data

```
# Clean NA value
# replace 000 with NA in DISTRICT_NUMBER and "Unknown" with NA in DISTRICT_NAME
FSA <- FSA %>% mutate(DISTRICT_NUMBER = ifelse(is.na(DISTRICT_NUMBER), "000", DISTRICT_NUMBER))
FSA <- FSA %>% mutate(DISTRICT_NAME = ifelse(is.na(DISTRICT_NAME), "Unknown", DISTRICT_NAME))
```

```
sum(is.na(FSA))
```

```
## [1] 0
```

```
# Modify data type
FSA$NUMBER_EXPECTED_WRITERS = as.numeric(as.character(FSA$NUMBER_EXPECTED_WRITERS))
```

```
## Warning: NAs introduced by coercion
```

```
# Cleaning "Msk"(values are fewer than 10) values
# Drop columns that contains Msk value (or NA value after data type transformation) in the "NUMBER_EXPE
FSA_filtered <- FSA[complete.cases(FSA$NUMBER_EXPECTED_WRITERS), ]
```

```
# Counting the total number of "Msk"
sum(FSA_filtered == "Msk", na.rm = TRUE)
```

```
## [1] 55118
```

```
# Replace all the MSK with values
# Condition 1: when value in the "NUMBER_EXPECTED_WRITERS" column is between 10-49, replace all "Msk" v

selected_columns <- c("NUMBER_WRITERS","NUMBER_UNKNOWN", "NUMBER_EMERGING", "NUMBER_ONTRACK", "NUMBER_E

FSA_filtered <- FSA_filtered %>%
  mutate_at(vars(selected_columns), function(x) ifelse(FSA_filtered$NUMBER_EXPECTED_WRITERS < 50 & x ==
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(selected_columns)
##
##   # Now:
##   data %>% select(all_of(selected_columns))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Condition 2: when value in the "NUMBER_EXPECTED_WRITERS" column is greater than 50, replace all "Msk"

selected_columns <- c("NUMBER_WRITERS","NUMBER_UNKNOWN", "NUMBER_EMERGING", "NUMBER_ONTRACK", "NUMBER_E

FSA_filtered <- FSA_filtered %>%
  mutate_at(vars(selected_columns), function(x) ifelse(FSA_filtered$NUMBER_EXPECTED_WRITERS >= 50 & x =
```

```
sum(FSA_filtered == "Msk", na.rm = TRUE)
```

```
## [1] 0
```

```
# change data type for the "NUMBER_WRITERS", "NUMBER_UNKNOWN", "NUMBER_EMERGING", "NUMBER_ONTRACK", "NU
FSA_filtered <- FSA_filtered %>%
  mutate(GRADE = as.numeric(GRADE),
         NUMBER_WRITERS = as.numeric(NUMBER_WRITERS),
         NUMBER_UNKNOWN = as.numeric(NUMBER_UNKNOWN),
         NUMBER_EMERGING = as.numeric(NUMBER_EMERGING),
         NUMBER_ONTRACK = as.numeric(NUMBER_ONTRACK),
         NUMBER_EXTENDING = as.numeric(NUMBER_EXTENDING),
         SCORE = as.numeric(SCORE))
```

```
# Convert SCHOOL_YEAR to numeric
FSA_filtered$SCHOOL_YEAR <- as.character(FSA_filtered$SCHOOL_YEAR)
FSA_filtered$SCHOOL_YEAR <- as.numeric(substring(FSA_filtered$SCHOOL_YEAR, 1, 4))
```

```
summary(FSA_filtered)
```

```
##   SCHOOL_YEAR    DATA_LEVEL         PUBLIC_OR_INDEPENDENT DISTRICT_NUMBER
##  Min.   :2007   Length:24222       Length:24222          Length:24222
##  1st Qu.:2010   Class :character   Class :character      Class :character
##  Median :2014   Mode  :character   Mode  :character      Mode  :character
##  Mean   :2014
##  3rd Qu.:2017
##  Max.   :2020
##  DISTRICT_NAME      SUB_POPULATION         GRADE       FSA_SKILL_CODE
##  Length:24222       Length:24222       Min.   :4.000   Length:24222
##  Class :character   Class :character   1st Qu.:4.000   Class :character
##  Mode  :character   Mode  :character   Median :7.000   Mode  :character
##                                        Mean   :5.506
##                                        3rd Qu.:7.000
##                                        Max.   :7.000
##  NUMBER_EXPECTED_WRITERS NUMBER_WRITERS  NUMBER_UNKNOWN    NUMBER_EMERGING
##  Min.   :   10           Min.   :    5   Min.   :    5.0   Min.   :    5.0
##  1st Qu.:   69           1st Qu.:   42   1st Qu.:   10.0   1st Qu.:   10.0
##  Median :  196           Median :  148   Median :   15.0   Median :   10.0
##  Mean   : 1370           Mean   : 1079   Mean   :  278.4   Mean   :  188.9
##  3rd Qu.:  531           3rd Qu.:  407   3rd Qu.:   82.0   3rd Qu.:   24.0
##  Max.   :50653           Max.   :44653   Max.   :19843.0   Max.   :12009.0
##  NUMBER_ONTRACK    NUMBER_EXTENDING     SCORE
##  Min.   :    5.0   Min.   :   5.0   Min.   :  0.000
##  1st Qu.:   10.0   1st Qu.:  10.0   1st Qu.:  6.732
##  Median :   74.0   Median :  10.0   Median :438.322
##  Mean   :  737.2   Mean   : 103.9   Mean   :309.626
##  3rd Qu.:  281.0   3rd Qu.:  10.0   3rd Qu.:475.933
##  Max.   :32840.0   Max.   :6548.0   Max.   :943.368
```

Part II – Expanding

Question1 : Calculate Yearly Growth

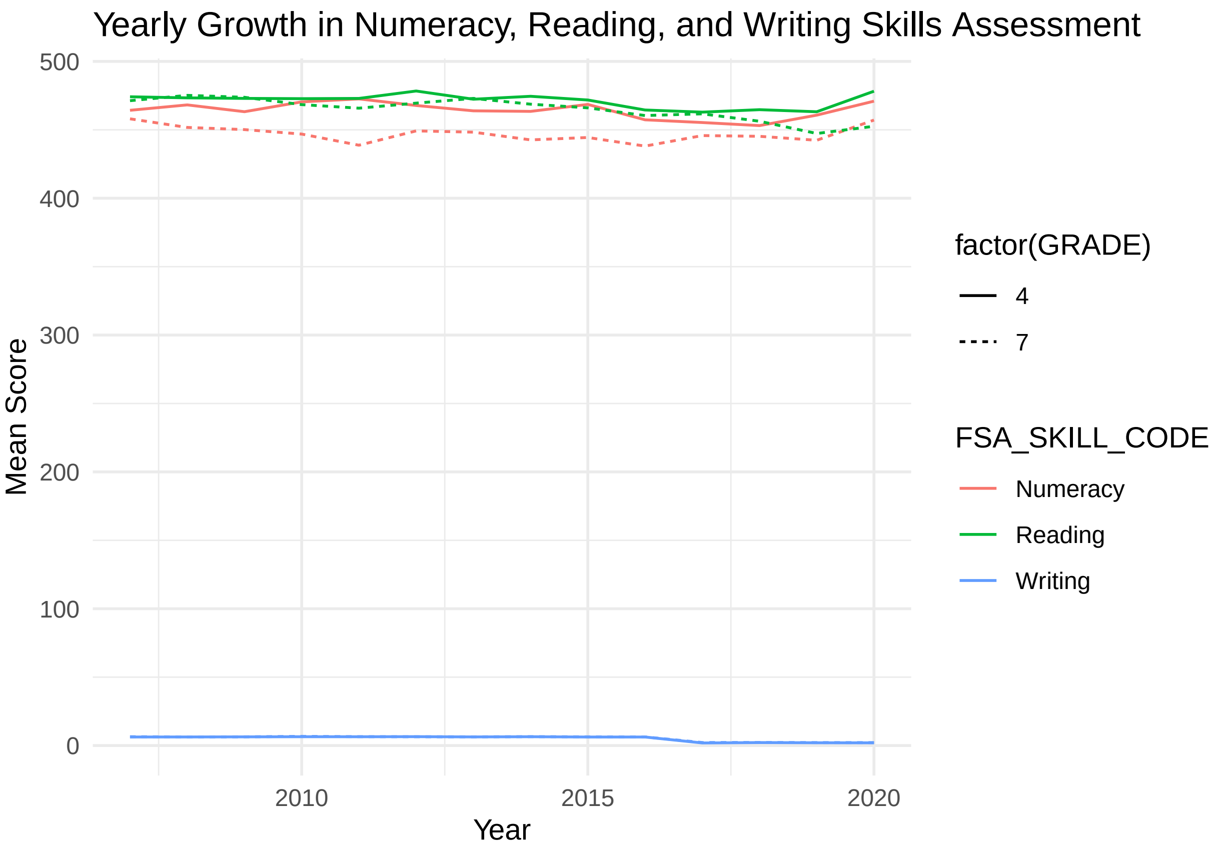
```
# Calculate the mean score for each year
mean_scores <- FSA_filtered %>%
  group_by(SCHOOL_YEAR, FSA_SKILL_CODE, GRADE) %>%
  summarise(mean_score = mean(SCORE, na.rm = TRUE))
```

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

```
print(mean_scores)
```

```
## # A tibble: 84 x 4
## # Groups:   SCHOOL_YEAR, FSA_SKILL_CODE [42]
##    SCHOOL_YEAR FSA_SKILL_CODE GRADE mean_score
##          <dbl> <chr>          <dbl>      <dbl>
##  1        2007 Numeracy           4     464.
##  2        2007 Numeracy           7     458.
##  3        2007 Reading            4     474.
##  4        2007 Reading            7     471.
##  5        2007 Writing            4       6.28
##  6        2007 Writing            7       6.47
##  7        2008 Numeracy           4     468.
##  8        2008 Numeracy           7     452.
##  9        2008 Reading            4     473.
## 10        2008 Reading            7     475.
## # i 74 more rows
```

```
ggplot(mean_scores, aes(x = SCHOOL_YEAR, y = mean_score, color = FSA_SKILL_CODE, linetype = factor(GRAD
  geom_line() +
  labs(title = "Yearly Growth in Numeracy, Reading, and Writing Skills Assessment",
       x = "Year",
       y = "Mean Score") +
  theme_minimal()
```



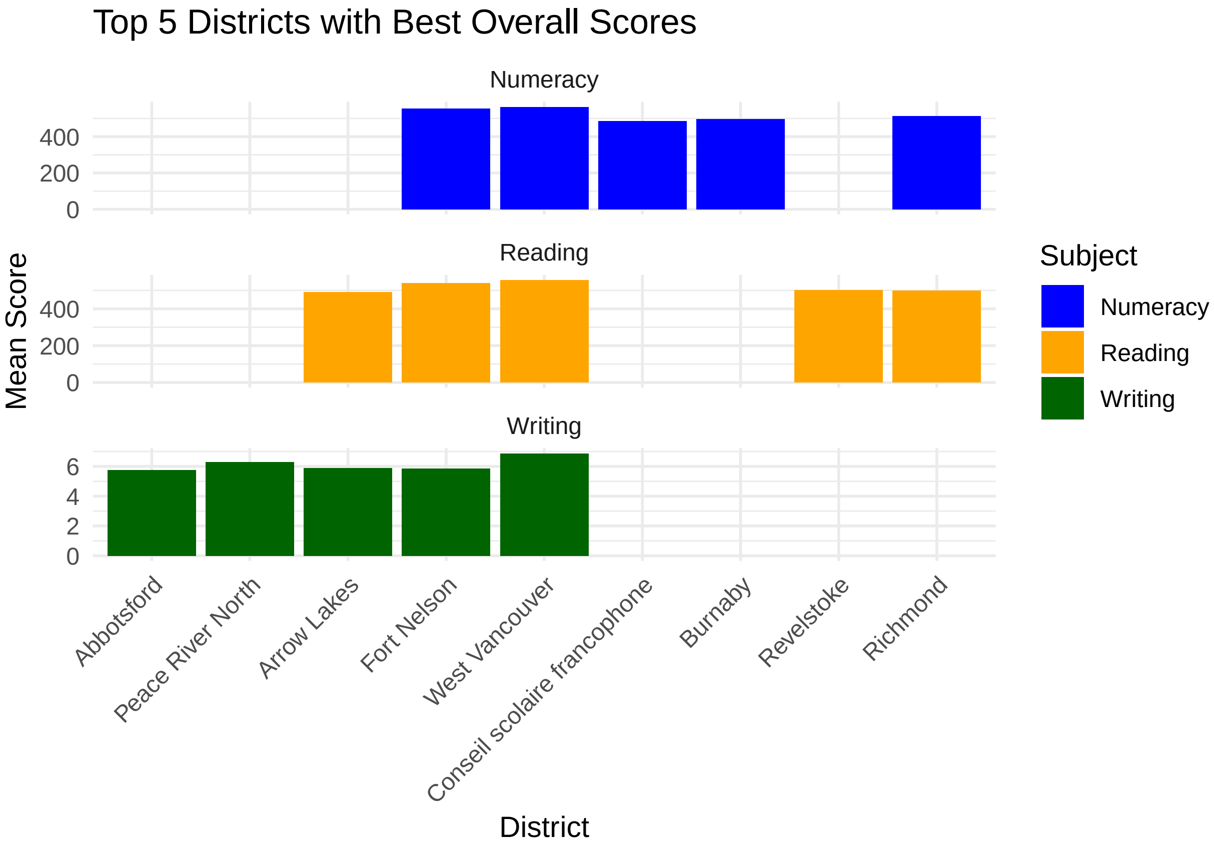
Question2 : Identify the top and worst 5 districts according to the overall scores from 2007 to 2021 for each subject

```
overall_mean_scores <- FSA_filtered %>%
  group_by(DISTRICT_NAME, FSA_SKILL_CODE) %>%
  summarise(mean_score = mean(SCORE, na.rm = TRUE))
```

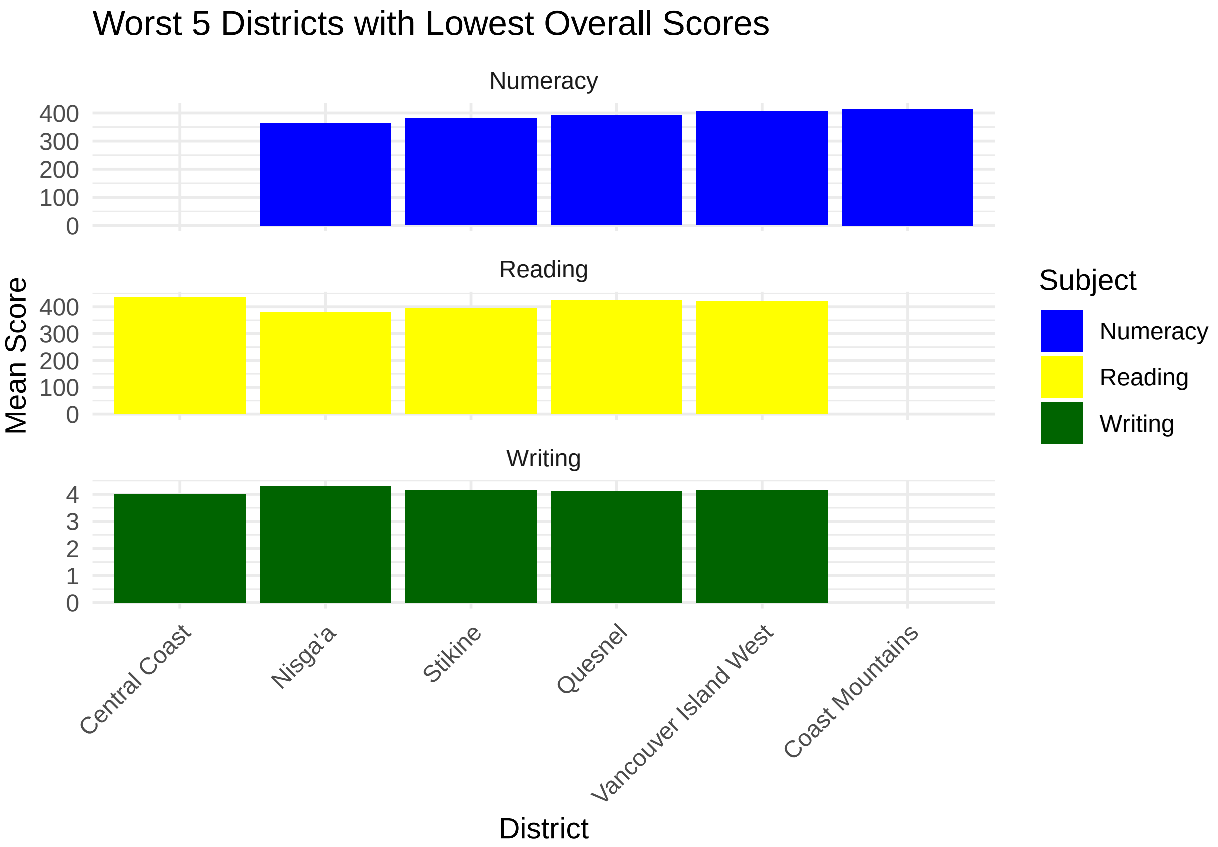
```
## `summarise()` has grouped output by 'DISTRICT_NAME'. You can override using the
## `.groups` argument.
```

```
# Identify the top 5 districts
top_districts <- overall_mean_scores %>%
  group_by(FSA_SKILL_CODE) %>%
  top_n(5, wt = mean_score) %>%
  ungroup()

# Identify the worst 5 districts
worst_districts <- overall_mean_scores %>%
  group_by(FSA_SKILL_CODE) %>%
  top_n(-5, wt = mean_score) %>%
  ungroup()
```

```
# Visualize the top 5 districts
ggplot(top_districts, aes(x = reorder(DISTRICT_NAME, mean_score), y = mean_score, fill = FSA_SKILL_CODE
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top 5 Districts with Best Overall Scores",
       x = "District",
       y = "Mean Score",
       fill = "Subject") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("Numeracy" = "blue", "Reading" = "orange", "Writing" = "darkgreen")) +
  facet_wrap(~FSA_SKILL_CODE, scales = "free_y", ncol = 1)
```



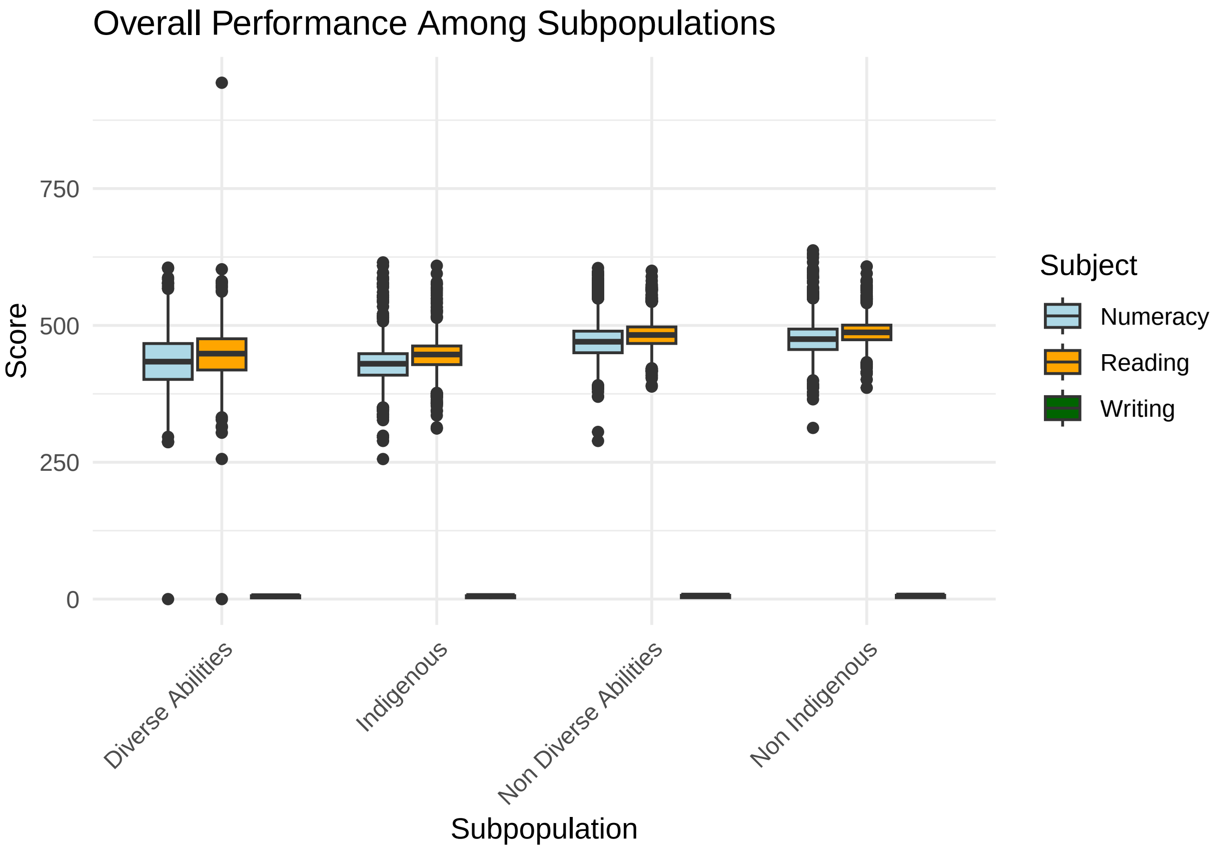
```
# Visualize the worst 5 districts

ggplot(worst_districts, aes(x = reorder(DISTRICT_NAME, mean_score), y = mean_score, fill = FSA_SKILL_CO
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Worst 5 Districts with Lowest Overall Scores",
       x = "District",
       y = "Mean Score",
       fill = "Subject") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("Numeracy" = "blue", "Reading" = "yellow", "Writing" = "darkgreen")) +
  facet_wrap(~FSA_SKILL_CODE, scales = "free_y", ncol = 1)
```



## Question3 : Compare the overall performance of different subpopulations

```
# Filter for the relevant columns and subpopulations
FSA_sub <- FSA_filtered %>%
  select(SUB_POPULATION, SCHOOL_YEAR, FSA_SKILL_CODE, SCORE) %>%
  filter(SUB_POPULATION %in% c('Indigenous', 'Diverse Abilities', 'Non Indigenous', 'Non Diverse Abiliti
```

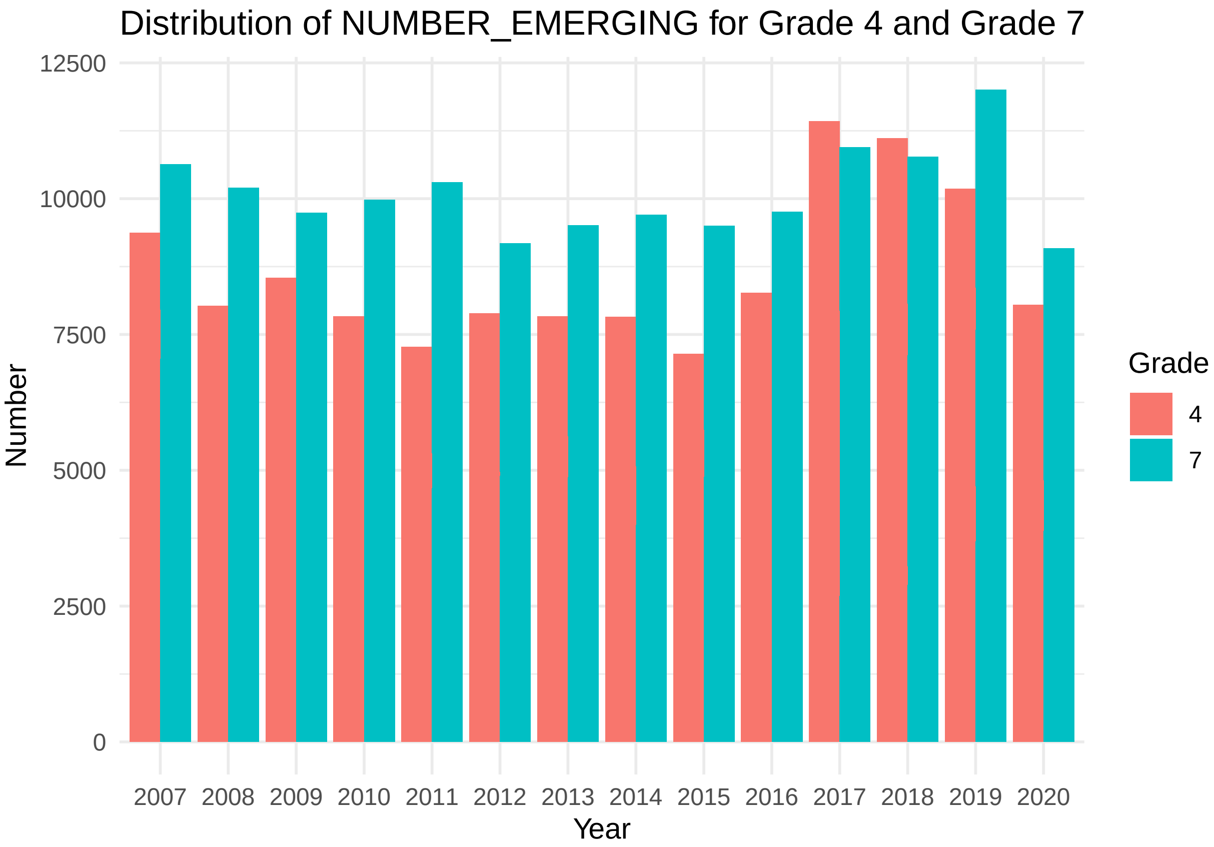
```
ggplot(FSA_sub, aes(x = SUB_POPULATION, y = SCORE, fill = FSA_SKILL_CODE)) +
  geom_boxplot() +
  labs(title = "Overall Performance Among Subpopulations",
       x = "Subpopulation",
       y = "Score",
       fill = "Subject") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("Numeracy" = "lightblue", "Reading" = "orange", "Writing" = "darkgreen")
```



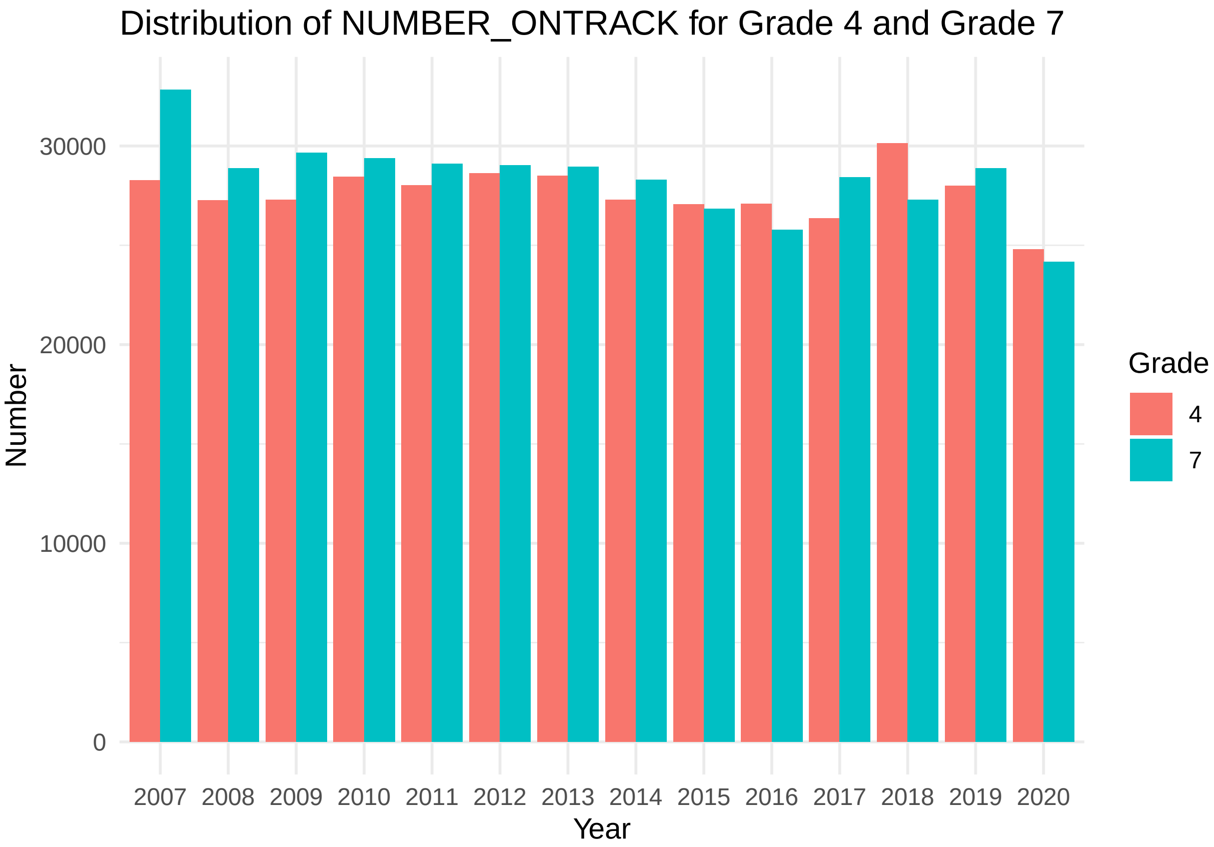
Question4 : How does the distribution change for student's performance on the test from 2017-2021?

```r
# Filter for the relevant columns
FSA_performance <- FSA_filtered %>%
  select(SCHOOL_YEAR, GRADE, NUMBER_EMERGING, NUMBER_ONTRACK, NUMBER_EXTENDING)
```

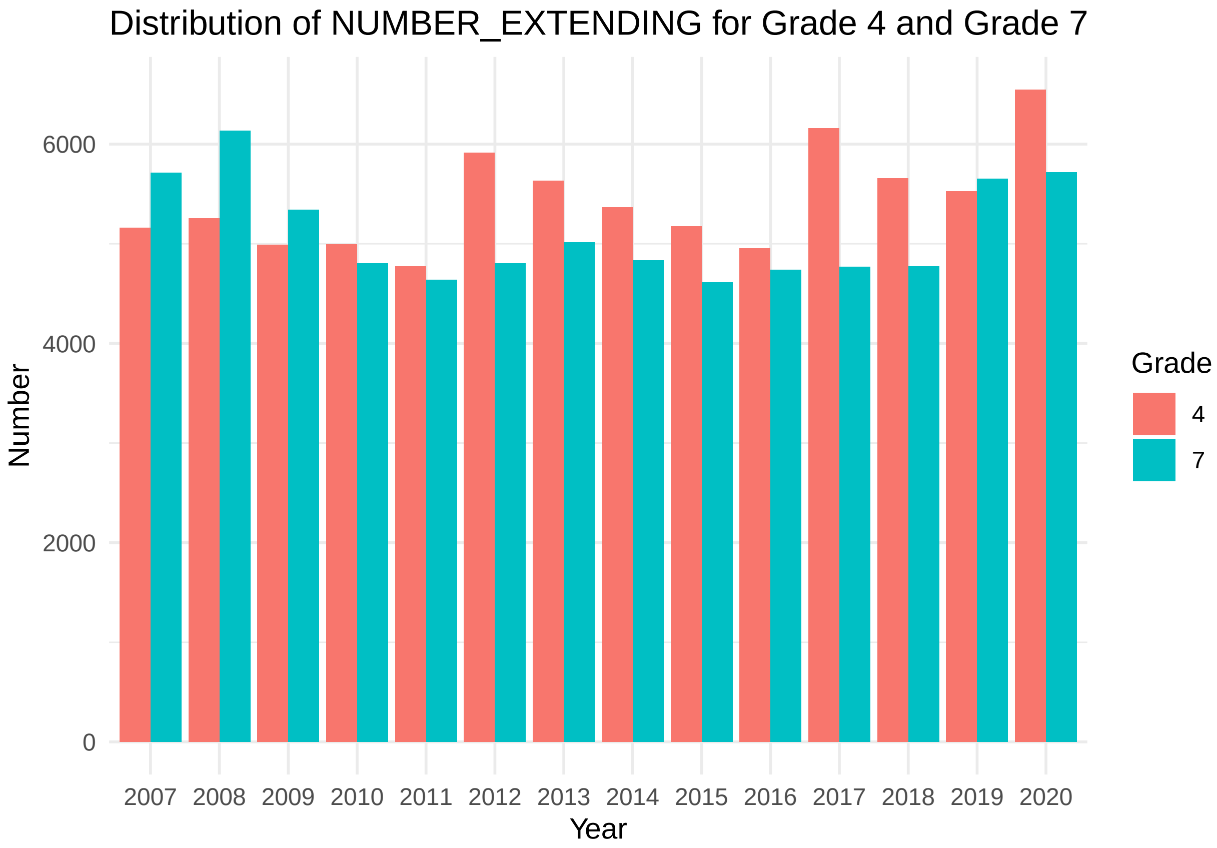
```r
# Visualize "NUMBER_EMERGING"
ggplot(FSA_performance, aes(x = factor(SCHOOL_YEAR), y = NUMBER_EMERGING, fill = factor(GRADE))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of NUMBER_EMERGING for Grade 4 and Grade 7",
       x = "Year",
       y = "Number",
       fill = "Grade") +
  theme_minimal()
```



```r
# Visualize "NUMBER_ONTRACK"
ggplot(FSA_performance, aes(x = factor(SCHOOL_YEAR), y = NUMBER_ONTRACK, fill = factor(GRADE))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of NUMBER_ONTRACK for Grade 4 and Grade 7",
       x = "Year",
       y = "Number",
       fill = "Grade") +
  theme_minimal()
```



```
# Visualize "NUMBER_EXTENDING"
ggplot(FSA_performance, aes(x = factor(SCHOOL_YEAR), y = NUMBER_EXTENDING, fill = factor(GRADE))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of NUMBER_EXTENDING for Grade 4 and Grade 7",
       x = "Year",
       y = "Number",
       fill = "Grade") +
  theme_minimal()
```


Distribution of NUMBER_EXTENDING for Grade 4 and Grade 7
Number
6000
4000
2000
0
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year
Grade
4
7