

607 Lab 6 Project2

Chi Hang(Philip) Cheung, Inna Yedzinovich

2024-10-11

```
options(repos = c(CRAN = "https://cran.r-project.org"))
library(tidyr)
suppressPackageStartupMessages(library(dplyr))
library(readr)
library(ggplot2)
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v forcats   1.0.0      v stringr   1.5.1
## v lubridate 1.9.3      v tibble   3.2.1
## v purrr     1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Dataset 1: SAT scores in 2010

```
url2<- "https://raw.githubusercontent.com/stormwhale/data-mines/refs/heads/main/SAT__College_Board__2010"
df2<- read.csv(url2)
view(df2)
colnames(df2)
```

```
## [1] "DBN" "School.Name" "Number.of.Test.Takers"
## [4] "Critical.Reading.Mean" "Mathematics.Mean" "Writing.Mean"
```

```
df2_tidy<- df2 %>%
  pivot_longer(cols = c('Critical.Reading.Mean', 'Mathematics.Mean', 'Writing.Mean'),
               names_to = 'Test_subjects',
               values_to = 'Average_scores',
               names_pattern = '(.*?)\\.Mean')
df2_tidy<-rename(df2_tidy,c('Number_of_test_takers'='Number.of.Test.Takers', 'School_name'='School.Name'))

#drop NA schools
df2_tidy<- df2_tidy %>%
  subset(!is.na(Number_of_test_takers))

#check if any NA still exist:
any(sum(is.na(df2_tidy)))
```

```
## [1] FALSE
```

```
view(df2_tidy)
```

Analyzing data

```
#schools with Critical Reading score>  
summary(df2_tidy) #Total number of school participated in the SAT = 386
```

```
##      DBN      School_name      Number_of_test_takers Test_subjects  
## Length:1158      Length:1158      Min.   :   7.0      Length:1158  
## Class :character      Class :character      1st Qu.: 35.0      Class :character  
## Mode  :character      Mode  :character      Median : 54.0      Mode  :character  
##                                           Mean   : 103.7  
##                                           3rd Qu.: 93.0  
##                                           Max.   :1047.0  
## Average_scores  
## Min.   :281  
## 1st Qu.:369  
## Median :389  
## Mean   :405  
## 3rd Qu.:421  
## Max.   :735
```

```
#To get the top 10 schools ranked by total SAT scores:  
top_total<-df2_tidy %>%  
  group_by(School_name) %>%  
  summarise(total_SAT_score= sum(Average_scores)) %>%  
  slice_max(total_SAT_score, n= 10)  
  
print(top_total)
```

```
## # A tibble: 10 x 2  
##   School_name      total_SAT_score  
##   <chr>          <int>  
## 1 "STUYVESANT HIGH SCHOOL "      2087  
## 2 "BRONX HIGH SCHOOL OF SCIENCE " 1960  
## 3 "STATEN ISLAND TECHNICAL HIGH SCHOOL " 1928  
## 4 "Townsend Harris High School at Queens College " 1923  
## 5 "HS of American Studies at Lehman College " 1884  
## 6 "QUEENS HS FOR SCIENCE YORK COL " 1875  
## 7 "BARD HIGH SCHOOL EARLY COLLEGE " 1868  
## 8 "BROOKLYN TECHNICAL HIGH SCHOOL " 1821  
## 9 "High School For Math Science Engineering City Coll " 1794  
## 10 "New York City Laboratory School Collab Studies " 1725
```

Top 10 schools SAT SCOREs break down by each test subject

```

top_sub <- df2_tidy %>%
  filter(School_name %in% top_total$School_name)
#To get the mean for each test subject from the top 10 schools:
top_sub_mean<- top_sub %>%
  group_by(Test_subjects) %>%
  summarise(Mean_score = mean(Average_scores))
print(top_sub_mean)

```

```

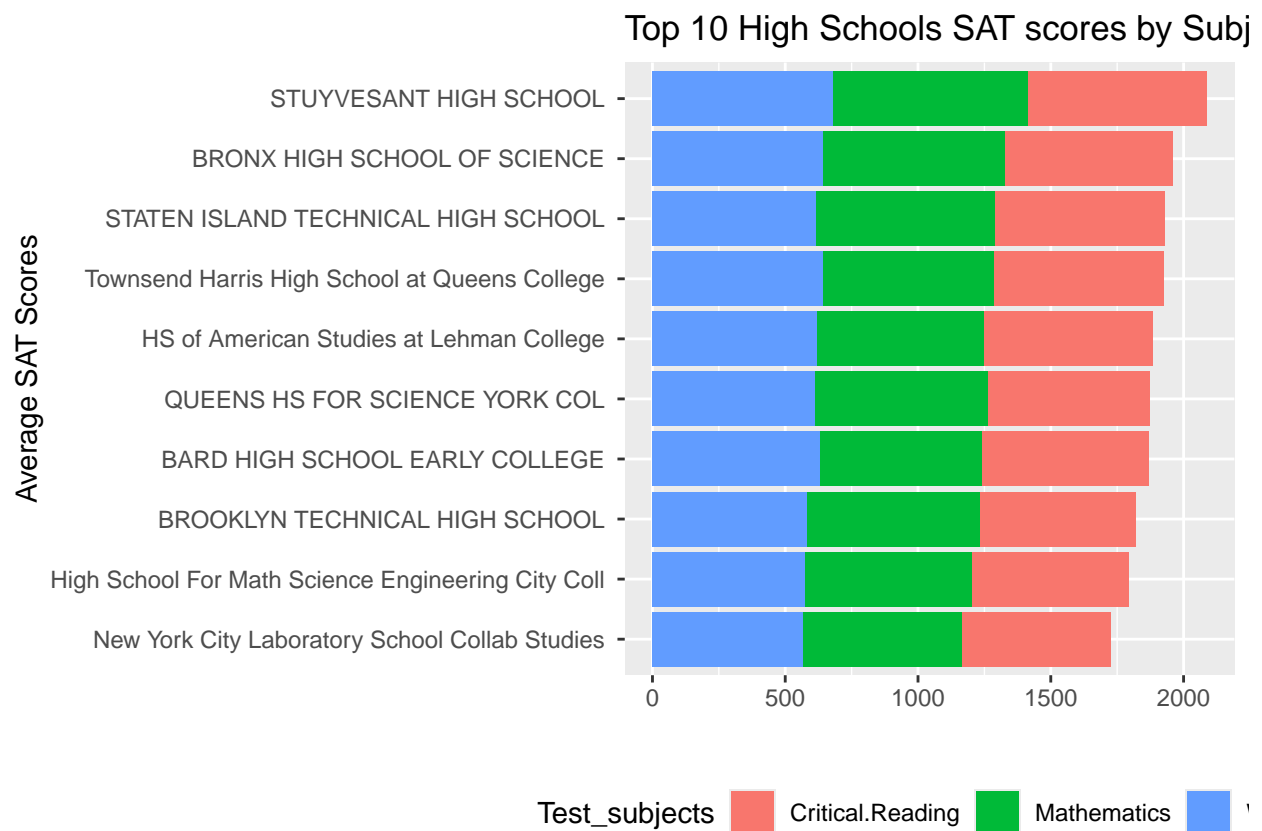
## # A tibble: 3 x 2
##   Test_subjects    Mean_score
##   <chr>          <dbl>
## 1 Critical.Reading    620
## 2 Mathematics        650.
## 3 Writing            616.

```

```

ggplot(top_sub, aes(x=reorder(School_name, Average_scores), y= Average_scores, fill=Test_subjects))+
  geom_bar(stat='identity') +
  coord_flip() +
  labs(title = 'Top 10 High Schools SAT scores by Subject', y='', x = 'Average SAT Scores')+
  theme(legend.position = 'bottom')

```



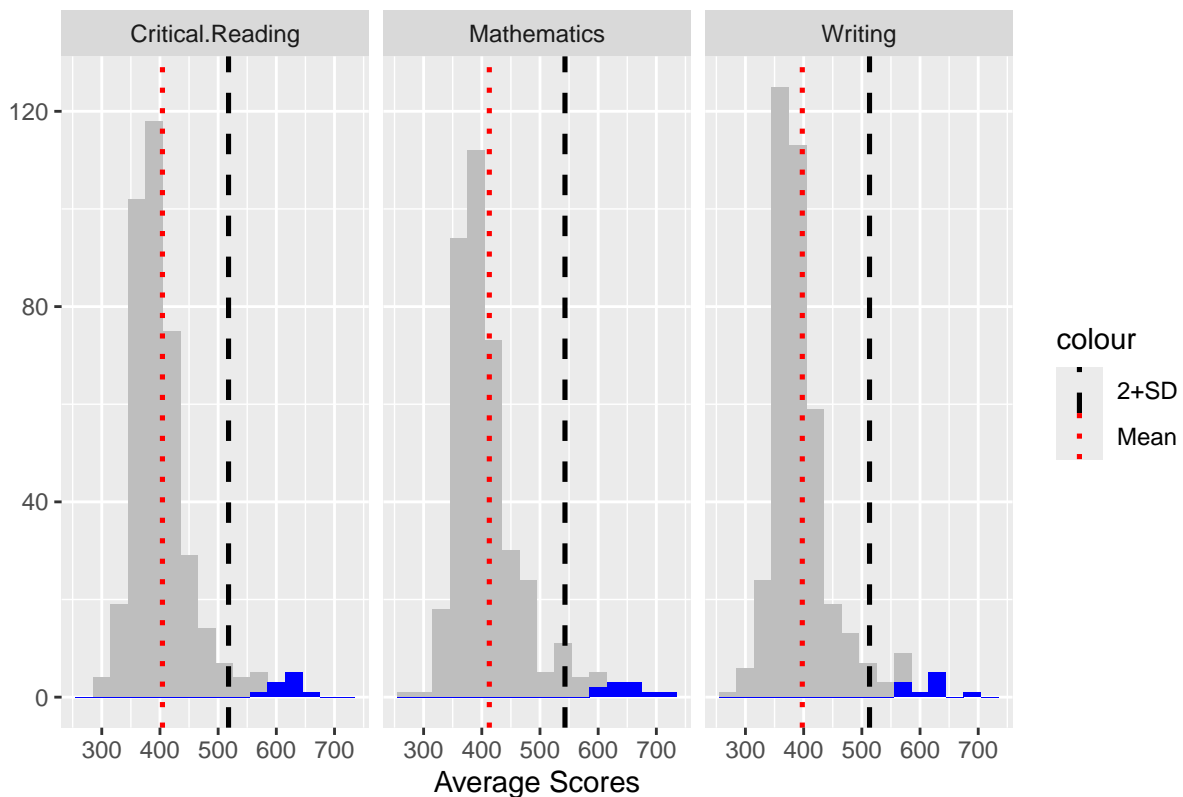
To look at how far ahead the top 10 schools from the mean scores of other schools:

```
#To get the mean and standard deviation from all HS that took the SAT:
tot_stat<- df2_tidy %>%
  group_by(Test_subjects) %>%
  summarise(Mean_score= mean(Average_scores), SD=sd(Average_scores), '1+SD'=Mean_score+SD, '2+SD'=Mean_score+2*SD)

tot_stat_mean<- tot_stat$Mean_score
tot_stat_2sd<- tot_stat$`2+SD`

ggplot() +
  geom_histogram(data= df2_tidy, aes(x=Average_scores), fill='grey', binwidth=30)+
  geom_histogram(data= top_sub, aes(x=Average_scores), fill='blue', binwidth=30)+
  facet_wrap(~Test_subjects)+
  labs(title = 'Top 10 high schools mean SAT score distribution by test subject',
       y='',
       x='Average Scores') +
  geom_vline(data= tot_stat, aes(xintercept= tot_stat_mean, color = 'Mean'), linetype='dotted', linewidth=1)+
  geom_vline(data= tot_stat, aes(xintercept= tot_stat_2sd, color = '2+SD'), linetype='dashed', linewidth=1)+
  scale_color_manual(values= c('Mean'='red', '2+SD'='black'))
```

Top 10 high schools mean SAT score distribution by test subject

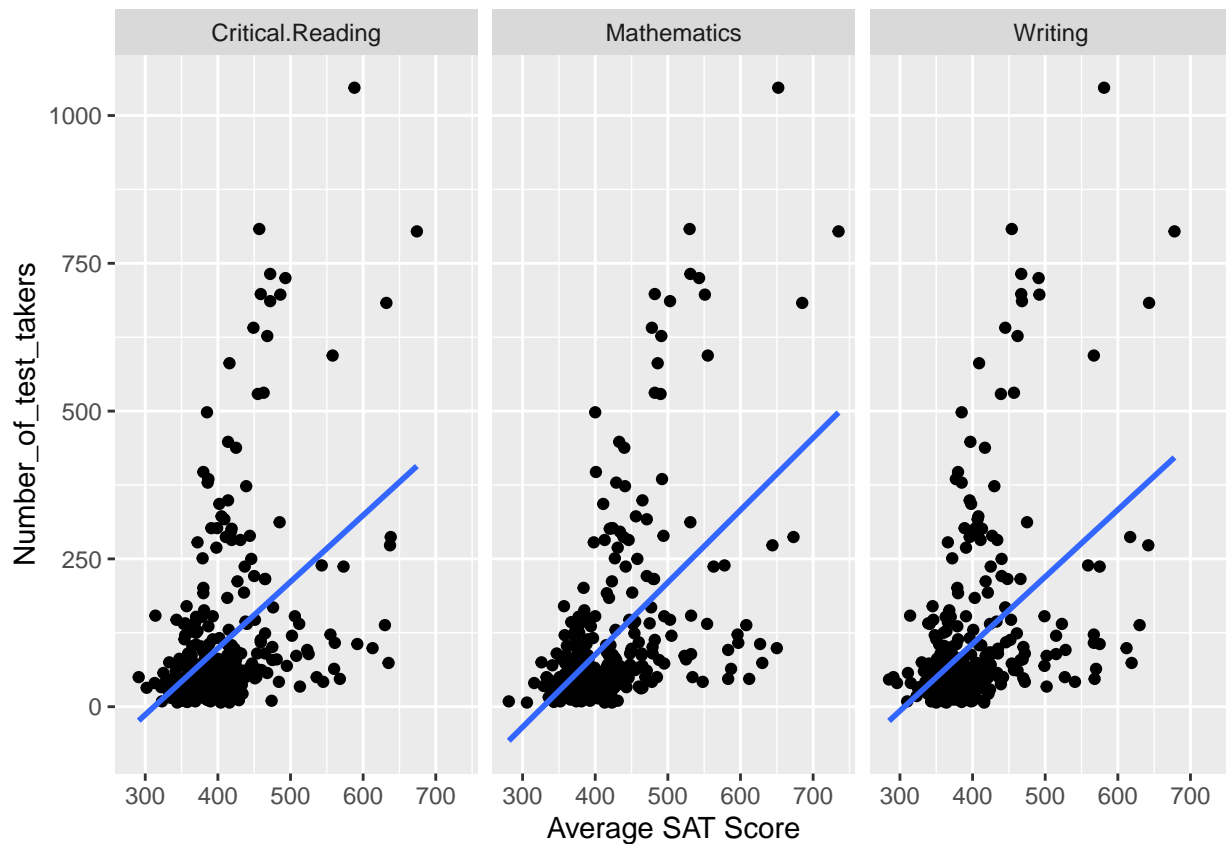


Conclusion: Top 10 High Schools are above 2 standard Deviation from the mean scores

Correlation between # of test takers and average SAT scores

```
ggplot(df2_tidy, aes(x= Average_scores, y=Number_of_test_takers))+  
  geom_point() +  
  facet_wrap(~Test_subjects) +  
  geom_smooth(method='lm', se=FALSE) +  
  labs(x= 'Average SAT Score')
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```



There is a positive correlation between number of test takers and average SAT scores. The higher number of people from a school participating in the SAT, the average score is generally higher than the mean value. This is observed in all three test subjects. =====

Dataset 2: Unity Data - MTA Daily Ridership

Data Overview The dataset contains daily ridership and traffic data for various transportation modes in New York City during March 2020. The columns include:

- Date: The date of the record.
- Subways: Total estimated ridership and percentage of comparable pre-pandemic day.
- Buses: Total estimated ridership and percentage of comparable pre-pandemic day.
- LIRR (Long Island Rail Road): Total estimated ridership and percentage of comparable pre-pandemic day.

- Metro-North: Total estimated ridership and percentage of comparable pre-pandemic day.
- Access-A-Ride: Total scheduled trips and percentage of comparable pre-pandemic day.
- Bridges and Tunnels: Total traffic and percentage of comparable pre-pandemic day.
- Staten Island Railway: Total estimated ridership and percentage of comparable pre-pandemic day.

Initial Analysis:

- There is a noticeable decline in ridership across all transportation modes as the month progresses. - The percentage of ridership compared to pre-pandemic levels shows a significant decline. - Access-A-Ride: This services maintained higher percentages of pre-pandemic levels compared to other modes, indicating continued demand for these services despite the pandemic - Traffic through bridges and tunnels also decreased but not as drastically as public transportation ridership. This could suggest a shift towards private vehicle useage during the pandemic.

```
url <- "https://raw.githubusercontent.com/Yedzinovich/Data-607/main/MTA_Daily_Ridership_Data.csv"
mta_data <- read_csv(url)
```

```
## Rows: 1671 Columns: 15
## -- Column specification -----
## Delimiter: ","
## chr (1): Date
## dbl (14): Subways: Total Estimated Ridership, Subways: % of Comparable Pre-P...
##
## 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.
```

```
mta_data <- mta_data %>% mutate(Date = as.Date(Date, format = "%m/%d/%y"))

mta_data_long <- mta_data %>% pivot_longer(cols = -Date, names_to = "Metric", values_to = "Value")
head(mta_data_long)
```

```
## # A tibble: 6 x 3
##   Date      Metric      Value
##   <date>    <chr>      <dbl>
## 1 2020-03-01 Subways: Total Estimated Ridership 2212965
## 2 2020-03-01 Subways: % of Comparable Pre-Pandemic Day 97
## 3 2020-03-01 Buses: Total Estimated Ridership 984908
## 4 2020-03-01 Buses: % of Comparable Pre-Pandemic Day 99
## 5 2020-03-01 LIRR: Total Estimated Ridership 86790
## 6 2020-03-01 LIRR: % of Comparable Pre-Pandemic Day 100
```

```
mta_data_long <- mta_data_long %>% separate(Metric, into = c("Transport_Mode", "Metric_Type"), sep = ":")
head(mta_data_long)
```

```
## # A tibble: 6 x 4
##   Date      Transport_Mode Metric_Type      Value
##   <date>    <chr>      <chr>      <dbl>
## 1 2020-03-01 Subways      Total Estimated Ridership 2212965
## 2 2020-03-01 Subways      % of Comparable Pre-Pandemic Day 97
## 3 2020-03-01 Buses      Total Estimated Ridership 984908
## 4 2020-03-01 Buses      % of Comparable Pre-Pandemic Day 99
## 5 2020-03-01 LIRR      Total Estimated Ridership 86790
## 6 2020-03-01 LIRR      % of Comparable Pre-Pandemic Day 100
```

Now that we have the data in a long format, we can extract more comprehensive insights from it. Long format can help us to perform a variety of analyses that are more flexible and insightful compared to the original wide format.

***What do know: - March 11, 2020, marks the start of the federal COVID-19 PHE declaration. - May 11, 2023, marks the end of the federal COVID-19 PHE declaration. Source:https://archive.cdc.gov/www_cdc.gov/coronavirus/2019-ncov/your-health/end-of-phe.html#:~:text=The%20federal%20COVID%2D19%20PHE,and%20testin

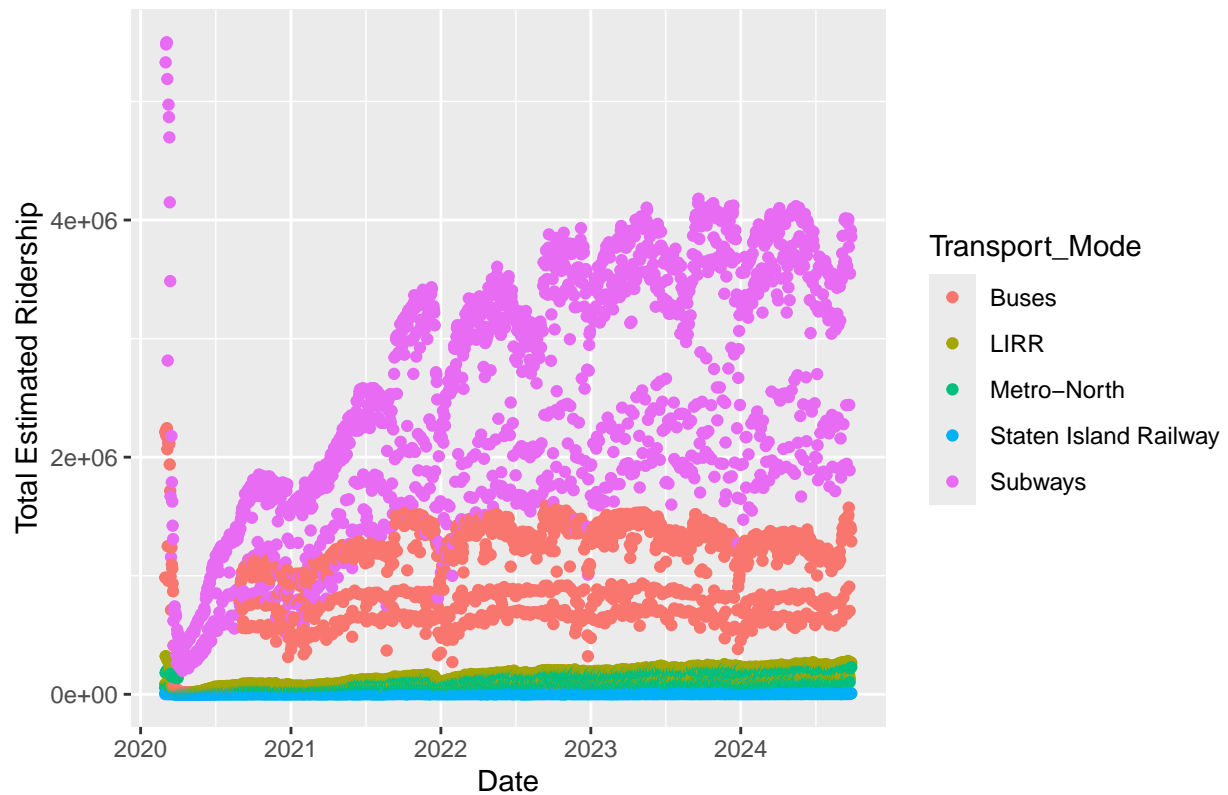
```
# Analysis #1
avg_ridership <- mta_data_long %>%
  filter(Metric_Type == "Total Estimated Ridership") %>%
  group_by(Transport_Mode) %>%
  summarize(Average_Ridership = mean(Value, na.rm = TRUE))

print(avg_ridership)
```

```
## # A tibble: 5 x 2
##   Transport_Mode      Average_Ridership
##   <chr>              <dbl>
## 1 Buses              1000673.
## 2 LIRR               134099.
## 3 Metro-North       113089.
## 4 Staten Island Railway 4382.
## 5 Subways           2482768.
```

```
# Analysis #2
ggplot(mta_data_long %>% filter(Metric_Type == "Total Estimated Ridership"), aes(x = Date, y = Value, c
  geom_point() +
  labs(title = "Public Transportation Ridership Trends Over Time", x = "Date", y = "Total Estimated Ridership")
```

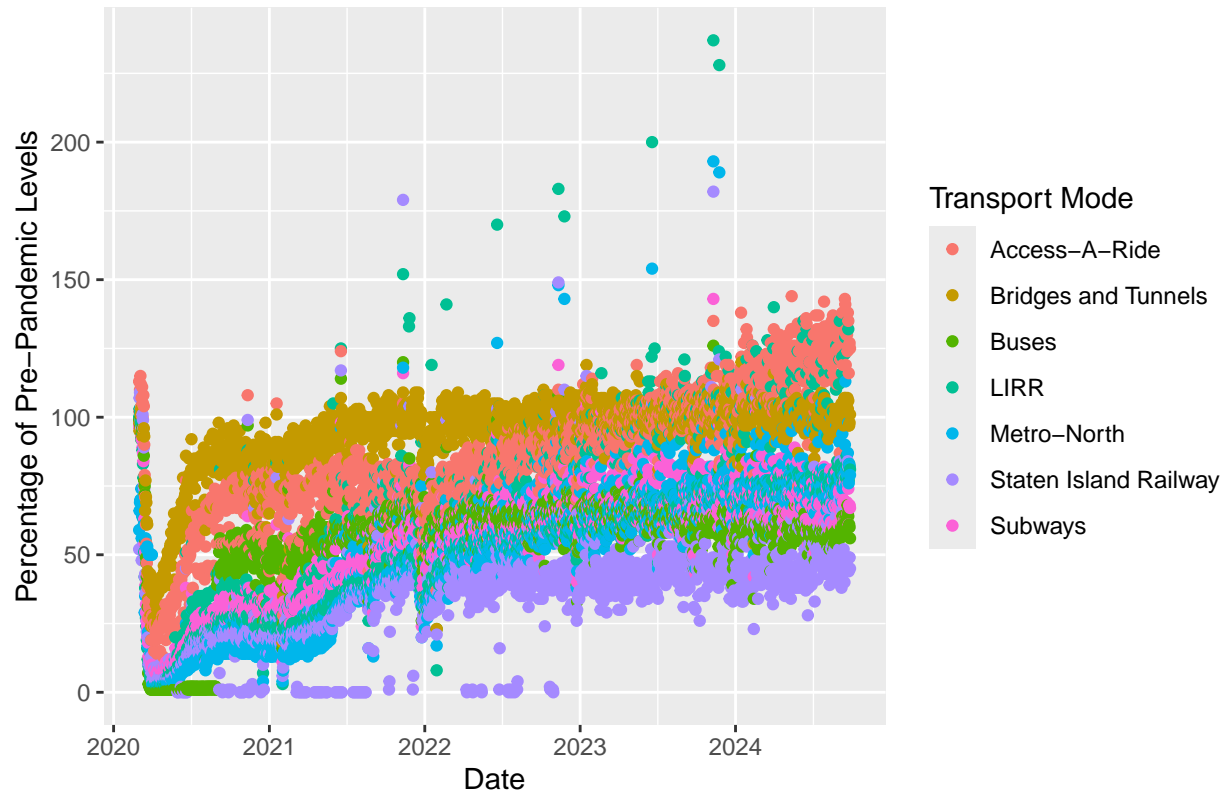
Public Transportation Ridership Trends Over Time



```
# Analysis #3
percentage_mta_data <- mta_data_long %>%
  filter(grepl("% of Comparable Pre-Pandemic Day", Metric_Type))

ggplot(permission_mta_data, aes(x = Date, y = Value, color = Transport_Mode)) +
  geom_point() +
  labs(title = "Percentage of Pre-Pandemic Levels Over Time",
       x = "Date",
       y = "Percentage of Pre-Pandemic Levels",
       color = "Transport Mode")
```

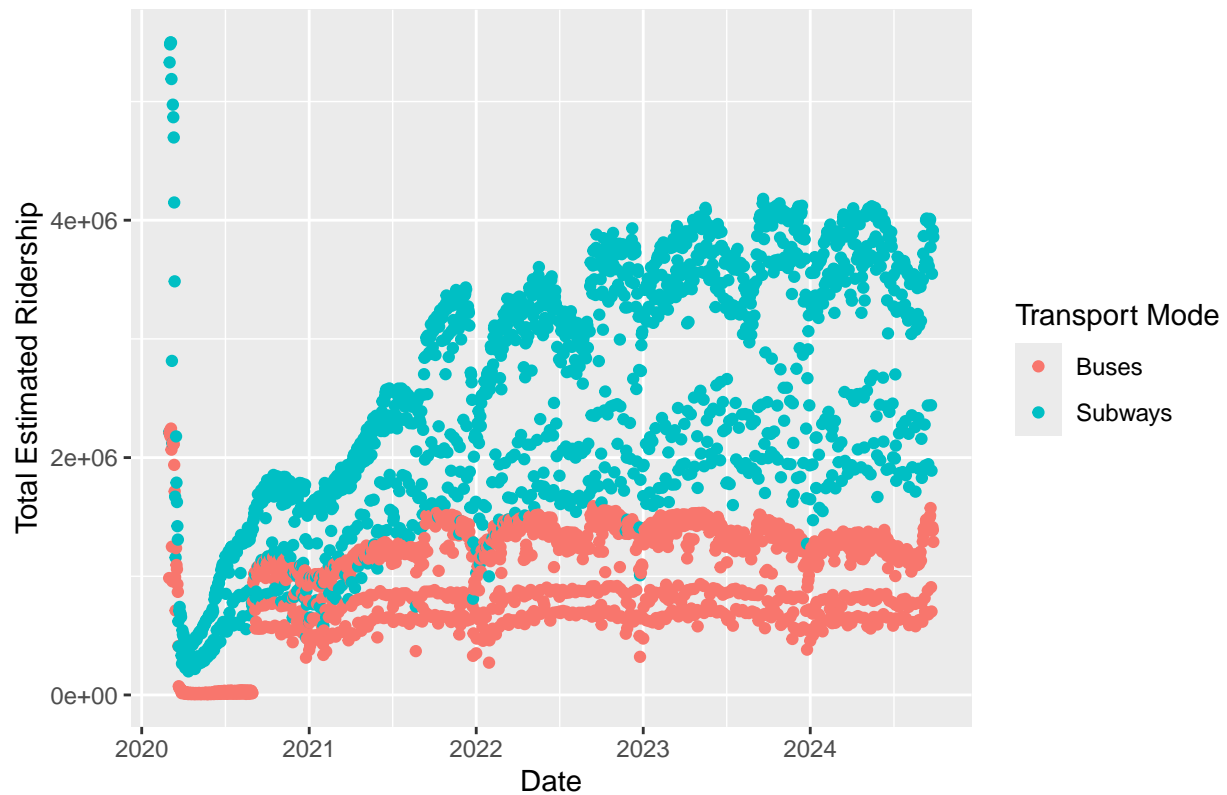

Percentage of Pre-Pandemic Levels Over Time



```
# Analysis 4
pt_mta_ridership_data <- mta_data_long %>%
  filter(Metric_Type == "Total Estimated Ridership")

ggplot(pt_mta_ridership_data %>% filter(Transport_Mode %in% c("Subways", "Buses")),
  aes(x = Date, y = Value, color = Transport_Mode)) +
  geom_point() +
  labs(title = "Ridership Trends Over Time: Subways vs Buses",
    x = "Date",
    y = "Total Estimated Ridership",
    color = "Transport Mode")
```

Ridership Trends Over Time: Subways vs Buses



```
# Calculate the average ridership for each transport mode during the pandemic
avg_ridership_pandemic <- pt_mta_ridership_data %>%
  filter(Date >= as.Date("2020-03-01") & Date <= as.Date("2023-05-11")) %>%
  group_by(Transport_Mode) %>%
  summarize(Average_Ridership = mean(Value, na.rm = TRUE))

print(avg_ridership_pandemic)
```

```
## # A tibble: 5 x 2
##   Transport_Mode      Average_Ridership
##   <chr>                <dbl>
## 1 Buses                952938.
## 2 LIRR                108023.
## 3 Metro-North         88261.
## 4 Staten Island Railway 3890.
## 5 Subways            2185483.
```

```
# Analysis 5
pre_pandemic <- pt_mta_ridership_data %>%
  filter(Date >= as.Date("2020-03-01") & Date <= as.Date("2020-03-11"))
print(pre_pandemic)
```

```
## # A tibble: 55 x 4
##   Date      Transport_Mode      Metric_Type      Value
##   <date>    <chr>                <chr>          <dbl>
```

```
## 1 2020-03-01 Subways Total Estimated Ridership 2212965
## 2 2020-03-01 Buses Total Estimated Ridership 984908
## 3 2020-03-01 LIRR Total Estimated Ridership 86790
## 4 2020-03-01 Metro-North Total Estimated Ridership 55825
## 5 2020-03-01 Staten Island Railway Total Estimated Ridership 1636
## 6 2020-03-02 Subways Total Estimated Ridership 5329915
## 7 2020-03-02 Buses Total Estimated Ridership 2209066
## 8 2020-03-02 LIRR Total Estimated Ridership 321569
## 9 2020-03-02 Metro-North Total Estimated Ridership 180701
## 10 2020-03-02 Staten Island Railway Total Estimated Ridership 17140
## # i 45 more rows
```

```
post_pandemic <- pt_mta_ridership_data %>%
  filter(Date >= as.Date("2023-05-12") & Date <= as.Date("2024-10-10"))
print(post_pandemic)
```

```
## # A tibble: 2,520 x 4
##   Date      Transport_Mode Metric_Type      Value
##   <date>    <chr>          <chr>          <dbl>
## 1 2023-05-12 Subways      Total Estimated Ridership 3723192
## 2 2023-05-12 Buses      Total Estimated Ridership 1436385
## 3 2023-05-12 LIRR       Total Estimated Ridership 201367
## 4 2023-05-12 Metro-North Total Estimated Ridership 185027
## 5 2023-05-12 Staten Island Railway Total Estimated Ridership 6629
## 6 2023-05-13 Subways      Total Estimated Ridership 2487178
## 7 2023-05-13 Buses      Total Estimated Ridership 918257
## 8 2023-05-13 LIRR       Total Estimated Ridership 113810
## 9 2023-05-13 Metro-North Total Estimated Ridership 109940
## 10 2023-05-13 Staten Island Railway Total Estimated Ridership 1973
## # i 2,510 more rows
```

```
avg_pre_pandemic <- pre_pandemic %>%
  group_by(Transport_Mode) %>%
  summarize(Average_Ridership_Pre = mean(Value, na.rm = TRUE))
print(avg_pre_pandemic)
```

```
## # A tibble: 5 x 2
##   Transport_Mode Average_Ridership_Pre
##   <chr>          <dbl>
## 1 Buses      1860634.
## 2 LIRR      236933.
## 3 Metro-North 154002.
## 4 Staten Island Railway 12476.
## 5 Subways    4425676.
```

```
avg_post_pandemic <- post_pandemic %>%
  group_by(Transport_Mode) %>%
  summarize(Average_Ridership_Post = mean(Value, na.rm = TRUE))
print(avg_post_pandemic)
```

```
## # A tibble: 5 x 2
##   Transport_Mode Average_Ridership_Post
```

```
##      <chr>                                <dbl>
## 1 Buses                                1111202.
## 2 LIRR                                194476.
## 3 Metro-North                        170578.
## 4 Staten Island Railway              5523.
## 5 Subways                           3171124.
```

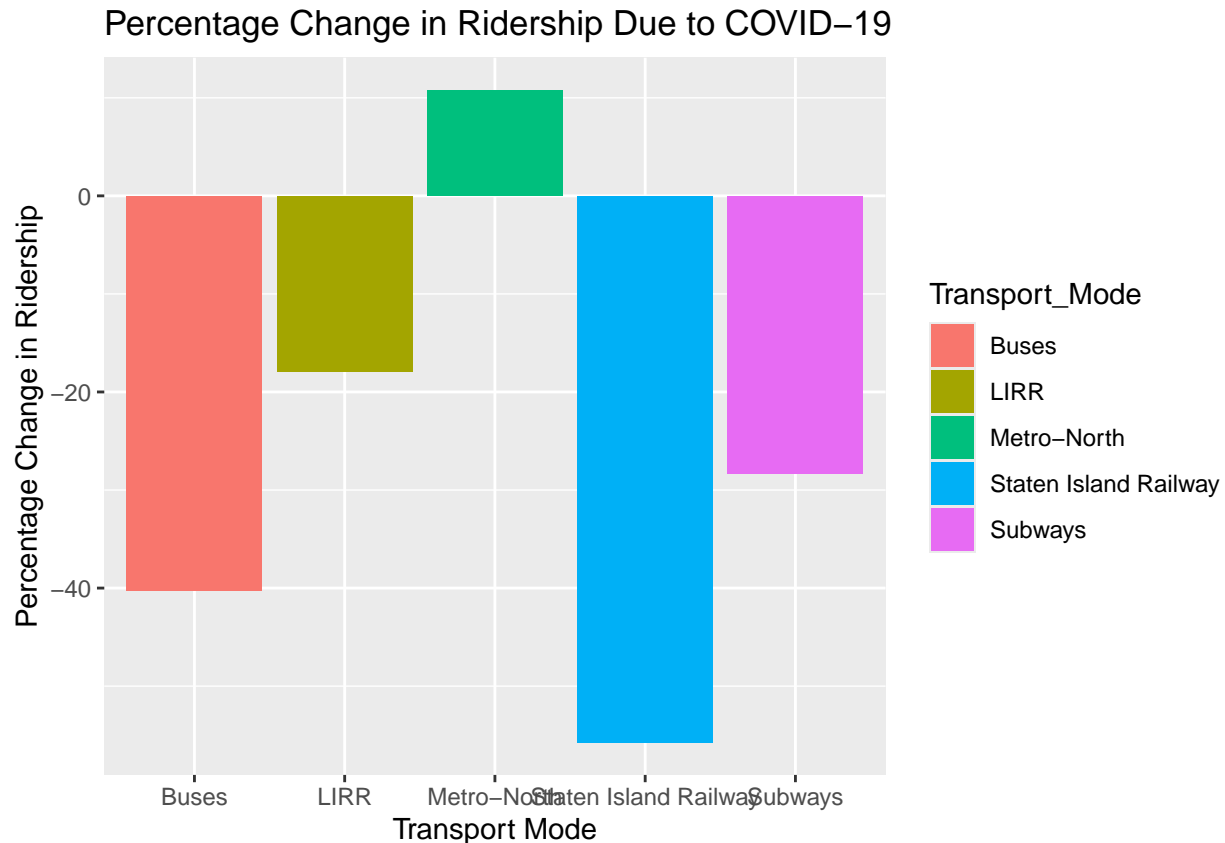
```
avg_ridership <- merge(avg_pre_pandemic, avg_post_pandemic, by = "Transport_Mode")
print(avg_ridership)
```

```
##      Transport_Mode Average_Ridership_Pre Average_Ridership_Post
## 1      Buses      1860634.45      1111202.327
## 2      LIRR      236932.91      194476.022
## 3  Metro-North      154001.82      170578.192
## 4 Staten Island Railway      12476.09      5523.302
## 5      Subways      4425676.18      3171124.308
```

```
avg_ridership <- avg_ridership %>%
  mutate(Percentage_Change = ((Average_Ridership_Post - Average_Ridership_Pre) / Average_Ridership_Pre))
print(avg_ridership)
```

```
##      Transport_Mode Average_Ridership_Pre Average_Ridership_Post
## 1      Buses      1860634.45      1111202.327
## 2      LIRR      236932.91      194476.022
## 3  Metro-North      154001.82      170578.192
## 4 Staten Island Railway      12476.09      5523.302
## 5      Subways      4425676.18      3171124.308
##      Percentage_Change
## 1      -40.27831
## 2      -17.91937
## 3       10.76375
## 4      -55.72891
## 5      -28.34712
```

```
ggplot(avg_ridership, aes(x = Transport_Mode, y = Percentage_Change, fill = Transport_Mode)) +
  geom_bar(stat = "identity") +
  labs(title = "Percentage Change in Ridership Due to COVID-19",
       x = "Transport Mode",
       y = "Percentage Change in Ridership")
```



Analysis: ##### 1- Analyze the average ridership for each transport mode. The data shows that subways have the most riders, with about 2.48 million people using them daily, much more than other types of transport. Buses come next, with around 1 million riders each day, making them very important. LIRR and Metro-North have fewer riders, with around 134,000 and 113,000 daily, since they serve commuters in specific regions. Staten Island Railway has the fewest riders, just over 4,000, likely because it covers a smaller area. Overall, subways and buses are the main ways people get around in the city.

2- Analyze how ridership for different transportation modes changes over time. This can help identify patterns, such as the impact of the COVID-19 pandemic on public transportation usage. As we can see, the chart shows public transportation ridership trends from 2020 to 2024, highlighting a sharp drop across all modes in early 2020 due to the COVID-19 pandemic, with subway ridership (in purple) experiencing the most significant decline. Ridership began recovering mid-2020, with buses showing a steadier recovery compared to the more volatile subway data. Regional transport modes like LIRR, Metro-North, and Staten Island Railway have consistently lower ridership. The chart reveals that, despite gradual recovery, ridership across all modes has not fully returned to pre-pandemic levels by 2024.

3- Analyze how the percentage of ridership compared to pre-pandemic levels changes over time for each transportation mode. The chart shows that public transportation usage dropped sharply during the pandemic but has been recovering at different rates across transport modes from 2020 to 2024. Bridges and tunnels saw the fastest recovery, exceeding 100% of pre-pandemic levels by 2021, indicating a shift towards car travel. Access-A-Ride and buses gradually returned to normal, nearing or slightly surpassing pre-pandemic levels by 2024. However, commuter services like the LIRR, Metro-North, and Staten Island Railway have been slower to recover, remaining below 100%, likely due to changes in work patterns. Subways are also recovering slowly, still below pre-pandemic levels by 2024.

4- Analyze the ridership trends between different transportation modes (subways vs buses) to see which modes were more resilient during the pandemic. The graph shows that both subway and bus ridership dropped sharply at the start of 2020 due to the pandemic. Subways saw a bigger drop than buses, but they have been recovering faster. By 2024, subway ridership has risen back to over 2 million, though it fluctuates more, while bus ridership has stayed steadier but remains below 2 million. Overall, subways have more riders than buses, but buses have a more stable number of users over time.

5- Analyze the impact of the COVID-19 pandemic on ridership by comparing pre-pandemic and post-pandemic data. The graph illustrates that most MTA transportation systems experienced a decline in ridership due to the COVID-19 pandemic, with Staten Island Railroad suffering the largest decrease, with over 50% fewer passengers. In contrast, Metro-North has seen an increase in ridership, likely due to people relocating from New York City during the pandemic and opting to commute using Metro-North.

Dataset 3: School diversity from 1994-2017

```
url3<- "https://raw.githubusercontent.com/stormwhale/data-mines/refs/heads/main/school%20divers.csv"
df3<-read.csv(url3)
view(df3)
```

Cleaning and tidying dataset. The racial groups are already represented in percentage for each school. NA values are assumed to be 0. All percentage will be rounded to the nearest tenth value.

```
df3_tidy<- df3 %>%
  mutate_all(~replace(., is.na(.), 0)) %>%
  pivot_longer(cols = AIAN:Multi,
               names_to = 'Ethnicity',
               values_to = 'Student_percentage') %>%
  mutate(Student_percentage=round(as.numeric(Student_percentage),1))

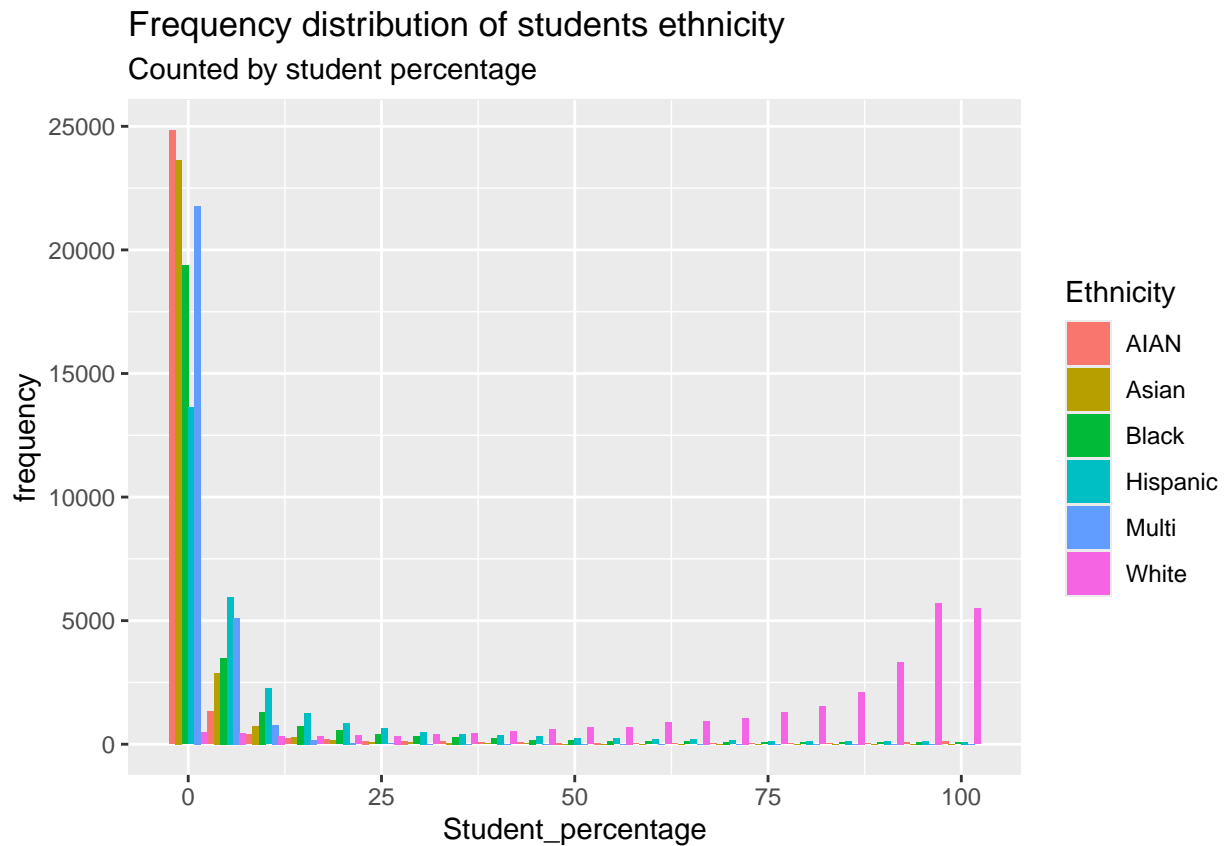
head(df3_tidy)
```

```
## # A tibble: 6 x 13
##       X.1      X  LEAID LEA_NAME      ST  d_Locale_Txt SCHOOL_YEAR Total diverse
##   <dbl> <dbl> <dbl> <chr>      <chr> <chr>      <chr>      <dbl> <chr>
## 1     1     1 100002 alabama youth~ AL     0          1994-1995    509 Diverse
## 2     1     1 100002 alabama youth~ AL     0          1994-1995    509 Diverse
## 3     1     1 100002 alabama youth~ AL     0          1994-1995    509 Diverse
## 4     1     1 100002 alabama youth~ AL     0          1994-1995    509 Diverse
## 5     1     1 100002 alabama youth~ AL     0          1994-1995    509 Diverse
## 6     1     1 100002 alabama youth~ AL     0          1994-1995    509 Diverse
## # i 4 more variables: variance <dbl>, int_group <chr>, Ethnicity <chr>,
## #   Student_percentage <dbl>
```

Data analysis:

1 - Overall distribution of ethnicity Asian and AIAN groups seem to be the most under integrated among different schools.

```
ggplot(df3_tidy, aes(x=Student_percentage, fill=Ethnicity)) +
  geom_histogram(binwidth = 5, position='dodge') +
  labs(y='frequency',
       title = 'Frequency distribution of students ethnicity',
       subtitle = 'Counted by student percentage')
```



In 2016-2017, which states have schools have integrated the most Asian and AIAN students?

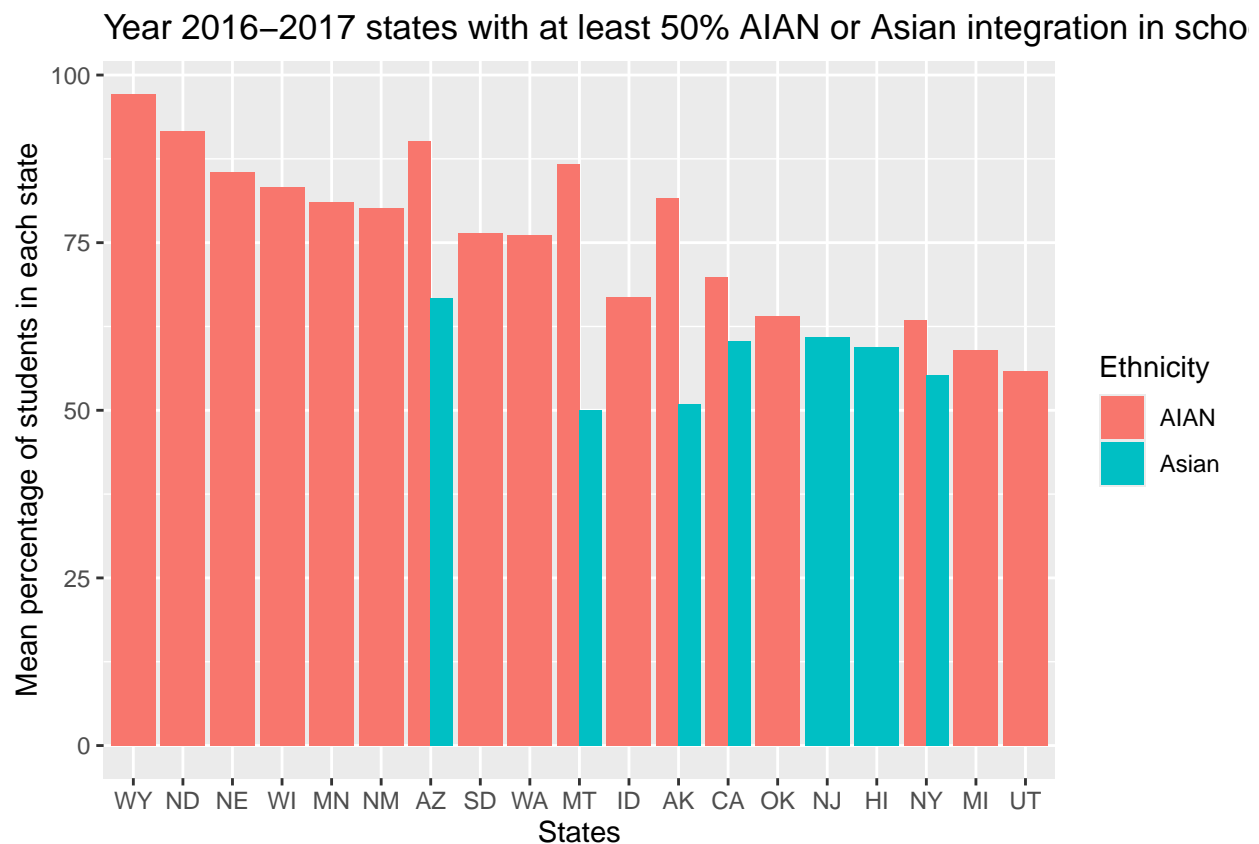
```
df3_2017<- df3_tidy %>%
  filter(SCHOOL_YEAR=='2016-2017') %>%
  filter((Ethnicity == 'Asian' | Ethnicity == 'AIAN') & Student_percentage >= 50) %>%
  group_by(ST, Ethnicity) %>%
  summarize(mean_percentage = mean(Student_percentage), .groups = 'drop') %>%
  arrange(desc(mean_percentage))

print(df3_2017)
```

```
## # A tibble: 24 x 3
##   ST      Ethnicity mean_percentage
##   <chr> <chr>          <dbl>
## 1 WY     AIAN             97.2
## 2 ND     AIAN             91.6
## 3 AZ     AIAN             90.2
## 4 MT     AIAN             86.7
## 5 NE     AIAN             85.5
## 6 WI     AIAN             83.3
```

```
## 7 AK      AIAN      81.7
## 8 MN      AIAN      81.0
## 9 NM      AIAN      80.1
## 10 SD     AIAN      76.4
## # i 14 more rows
```

```
ggplot(df3_2017, aes(x=reorder(ST, -mean_percentage), y=mean_percentage, fill= Ethnicity)) +
  geom_bar(stat='identity', position = 'dodge') +
  labs(title = 'Year 2016-2017 states with at least 50% AIAN or Asian integration in schools',
        x='States',
        y='Mean percentage of students in each state')
```

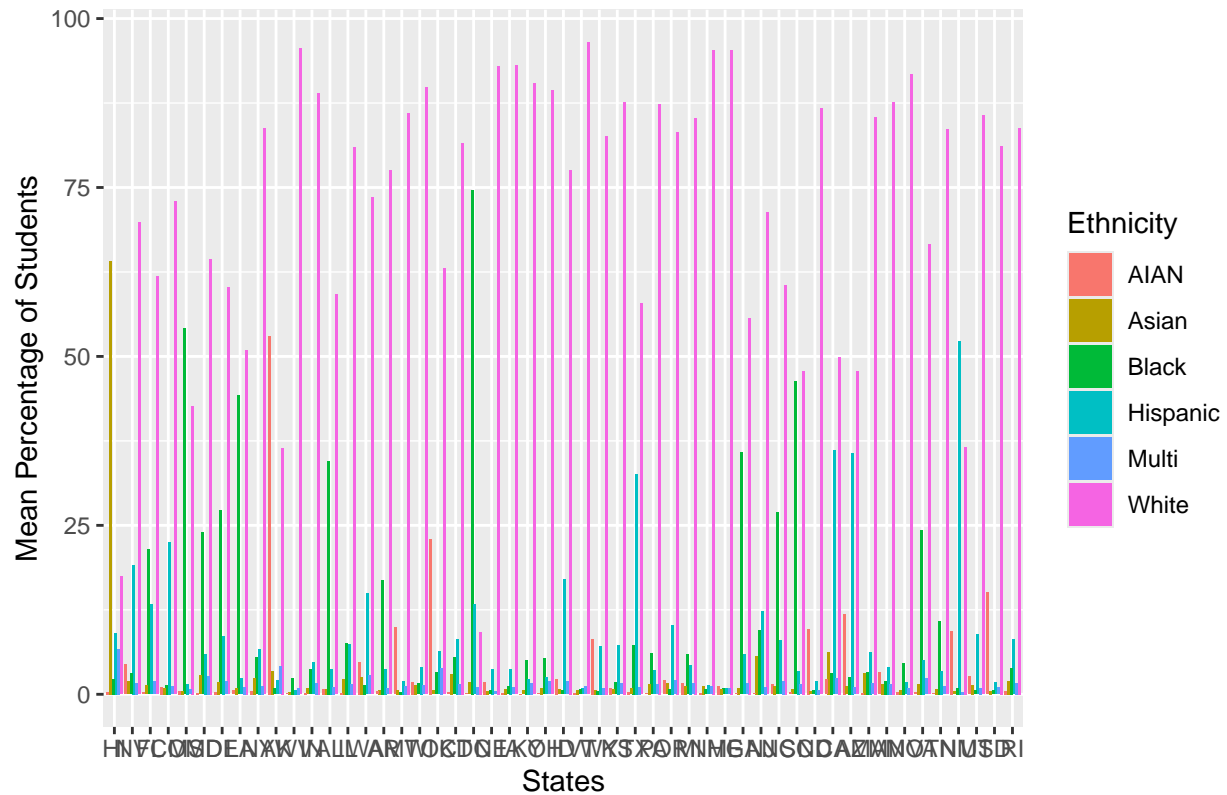


2 - Between State Comparison Compare the integration levels of different ethnic groups between states to see which states are more integrated.

```
df3_state_comparison <- df3_tidy %>%
  group_by(ST, Ethnicity) %>%
  summarize(mean_percentage = mean(Student_percentage), .groups = 'drop') %>%
  arrange(desc(mean_percentage))

ggplot(df3_state_comparison, aes(x=reorder(ST, -mean_percentage), y=mean_percentage, fill=Ethnicity)) +
  geom_bar(stat='identity', position='dodge') +
  labs(title = 'Comparison of Ethnic Integration Between States',
        x='States',
        y='Mean Percentage of Students')
```


Comparison of Ethnic Integration Between States



Analysis: #### Analysis 1: Overall distribution of ethnicity The graph depicts the mean percentage of students in U.S. states during the 2016-2017 school year, with at least 50% of either American Indian and Alaska Native or Asian students integrated into schools. Wyoming, North Dakota, and Nebraska have the highest proportions of AIAN students, approaching or at 100%. In contrast, states like Arizona, California, and New Jersey show substantial integration of Asian students. States like South Dakota, Washington, and Idaho present a more balanced distribution between AIAN and Asian student populations. Overall, the chart highlights significant regional differences in ethnic integration within school systems, particularly for AIAN and Asian students

Analysis 2: Between State Comparison The graph compares the ethnic integration of students in various states, showing the mean percentage of different racial groups in schools. White students make up the largest proportion of students in most states, often exceeding 50% and even approaching 100% in several. Hispanic, Black, and Asian students are represented at lower percentages across states, with some variation. AIAN students and students of multiple races have smaller but noticeable percentages in some states. The chart highlights significant racial differences in student populations across states, with White students being the most dominant group.