EDLD 652 Final Project Draft

Cassie Malcom Merly Klaas Havi Khurana

2/21/2022

Contents


```
knitr::opts_chunk$set(echo = TRUE,
                      warning = FALSE,
                      message = FALSE,
                      error = FALSE,
                      fig.width = 9,
                      fig.height = 9)
```

```
pacman::p_load("tidyverse","rio","here","janitor", "usmap","maps", "colorspace","geofacet","leaidr", "vi
```

RQ1

Student and Teacher ethnic distribution in K-12 public schools in the US

```r
#We used "district-membership-17-21.parquet" file. Being too big, we created a subset of this data and
dm <- read_parquet(here("district-membership-17-21.parquet"))

dm_s <- dm %>%
  select(GRADE,LEAID,LEA_NAME,RACE_ETHNICITY,
         SCHOOL_YEAR,SEX,ST,STATENAME, STUDENT_COUNT, YEAR) %>%
  filter(YEAR %in% c("2017","2018")) #some teacher-ethnicity data is for year 2017
rm(dm) #freeing space

unique(dm_s$GRADE)
#"Grade 6"           "Grade 7"           "Grade 8"
#[4] "Grade 9"           "Kindergarten"      "Not Specified"
#[7] "Pre-Kindergarten"  "Ungraded"          "No Category Codes"
#[10] "Grade 1"           "Grade 10"          "Grade 11"
#[13] "Grade 12"          "Grade 2"           "Grade 3"
#[16] "Grade 4"           "Grade 5"           "Adult Education"
#[19] "Grade 13"

#I'm making two subsets here. One which has only K12 student distribution by race and ethnincity, and a

#Student data for K12

#Let's remove the other grade classes.

dm_k12 <- subset(dm_s, grepl(("^G|^K"), GRADE))

unique(dm_k12$GRADE) #We still have Grade 13, let's remove that

dm_k12 <- dm_k12 %>%
    filter(GRADE != "Grade 13")

export(dm_k12, here("data", "dm_k12.rda"))

# Total students in district

#"No Category Codes" is in the GRADE, RACE/ETHNICITY, and SEX categories.
#This comprises the sum of all the other groups at the category level.
#We checked this compaing the two values; it almost all cases this was equal. (for 3 million)
#In a handful of cases (<50), this was not equal when no grade-wise student ethnincity data was availab
#We also want total students enrolled for each district irrespective of the grades.
#This information is coded in grade == "no category codes", race/ethnicity == "no category codes", and

dm_total <- dm_s %>%
  filter(YEAR == "2018",
         GRADE == "No Category Codes",
         RACE_ETHNICITY == "No Category Codes",
         SEX == "No Category Codes")

#weird that each district is occuring two times. let's just keep one.

dm_total <- dm_total %>%
  distinct()
```

```
length(unique(dm_total$LEAID)) #this doesn't match our dm_total dimensions.

dm_total <- dm_total %>%
  distinct() %>%
  group_by(LEAID) %>%
  mutate(n=n()) #some districts have two rows

#Some districts (44) have double reporting.
#Mostly in DC: On checking, one number points to 0 and another to a finite value.
#In all other cases (NV, OR, VT), both rows have very close values.
#Let's keep the higher of the two.

temp <- dm_total %>%
  filter(n ==2) %>%
  slice(which.max(STUDENT_COUNT))

#Let's join them

dm_total <- dm_total %>%
  filter(n == 1) %>%
  rbind(temp)

#Now each district has a unique row, and there are no inconsistencies.

#export it
export(dm_total, here("data","dm_total.rda"))
```

```
dm_k12 <- import(here("data","dm_k12.rda")) %>%
    clean_names()
```

```
#Let's pool student population by grades and gender into a single ethnic categories district membership

dm_sum <- dm_k12 %>%
    group_by(st, leaid, race_ethnicity, year, statename) %>%
    summarise(
        student = sum(student_count, na.rm = TRUE) #students belonging to one race
    ) %>%
    group_by(st, leaid, year) %>%
    mutate(
        total = sum(student, na.rm = TRUE),
        no_code = sum(if_else(race_ethnicity == "No Category Codes", student, 0)),
        total_reported = total - no_code,
        flag = ifelse(no_code == total_reported, TRUE, FALSE)
    )
```

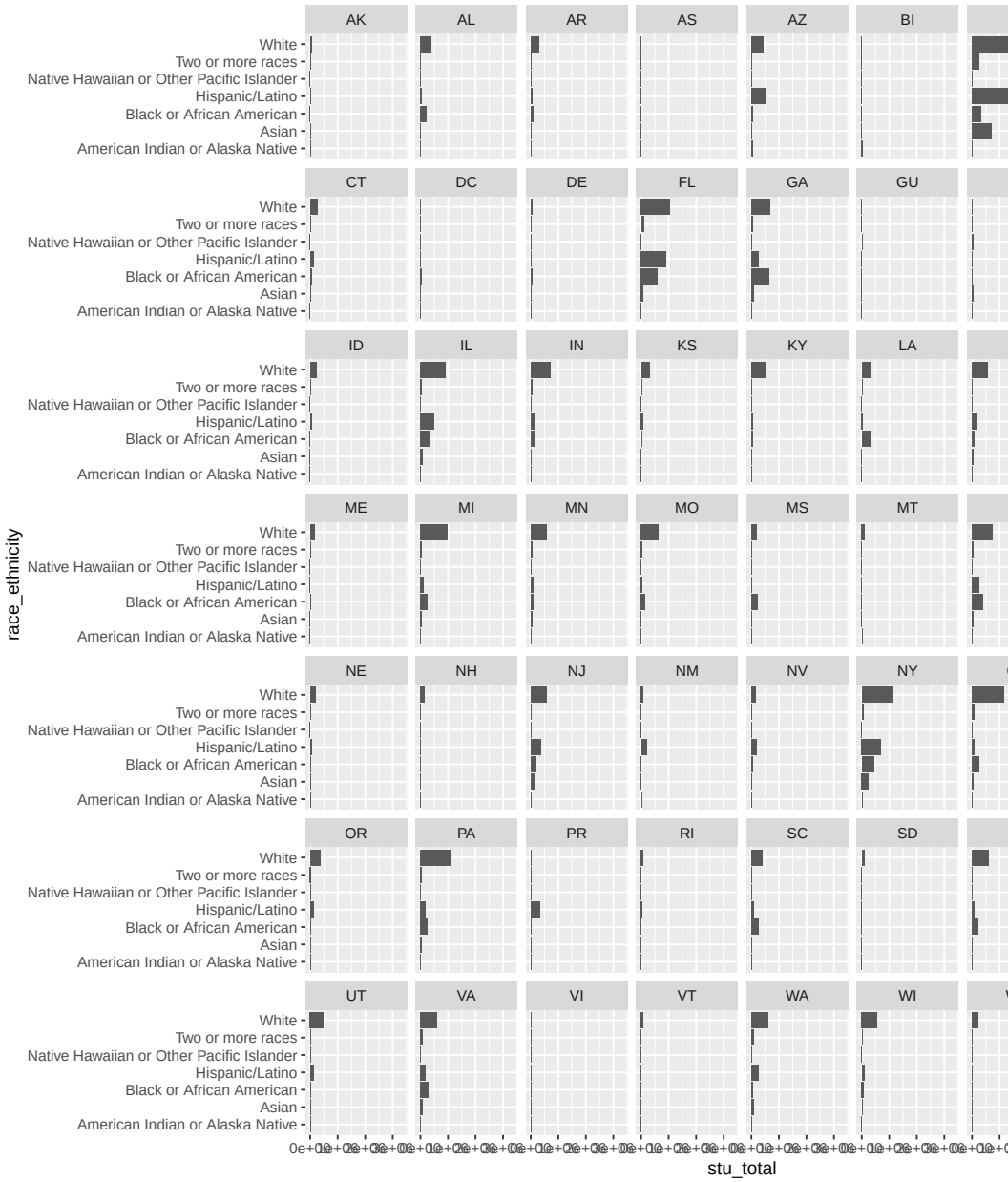
```
#quick check
dm_sum %>%
   group_by(flag) %>%
  summarise(
      n = n()
   )

#315612 times no_code = total_reported. Only 45 times this is not the case.
```

```
#This doesn't seem like a coincidence. It feels like students were double counted
#If this is true, total_reported would be the correct number of total students.

#Let's explore the FALSE situations
flag_f <- dm_sum %>%
    filter(flag == FALSE)

#this only happens in 5 districts (3 CA, 2 KS) for 2018 year where student by ethnicity data is not pre
#i.e., students in each sub-group is 0.

#Next, let's exclude these false situations, and find percentage of each ethnic group for 2018 and leai
```

```
dm18_long <- dm_sum %>%
    filter(year == "2018" & flag != "FALSE" & !race_ethnicity %in% c("No Category Codes", "Not Specifie
    mutate(
        percent_d = round((student/total_reported)*100, 3)
    ) %>%
    select(leaid, race_ethnicity, student, percent_d, total_reported, everything())

#data in wide format
dm18_wide <- dm18_long %>%
    pivot_wider(
        names_from = race_ethnicity,
        values_from = c(student,percent_d)
    )

#rm(dm_k12,dm_sum)
```

Prelim plots

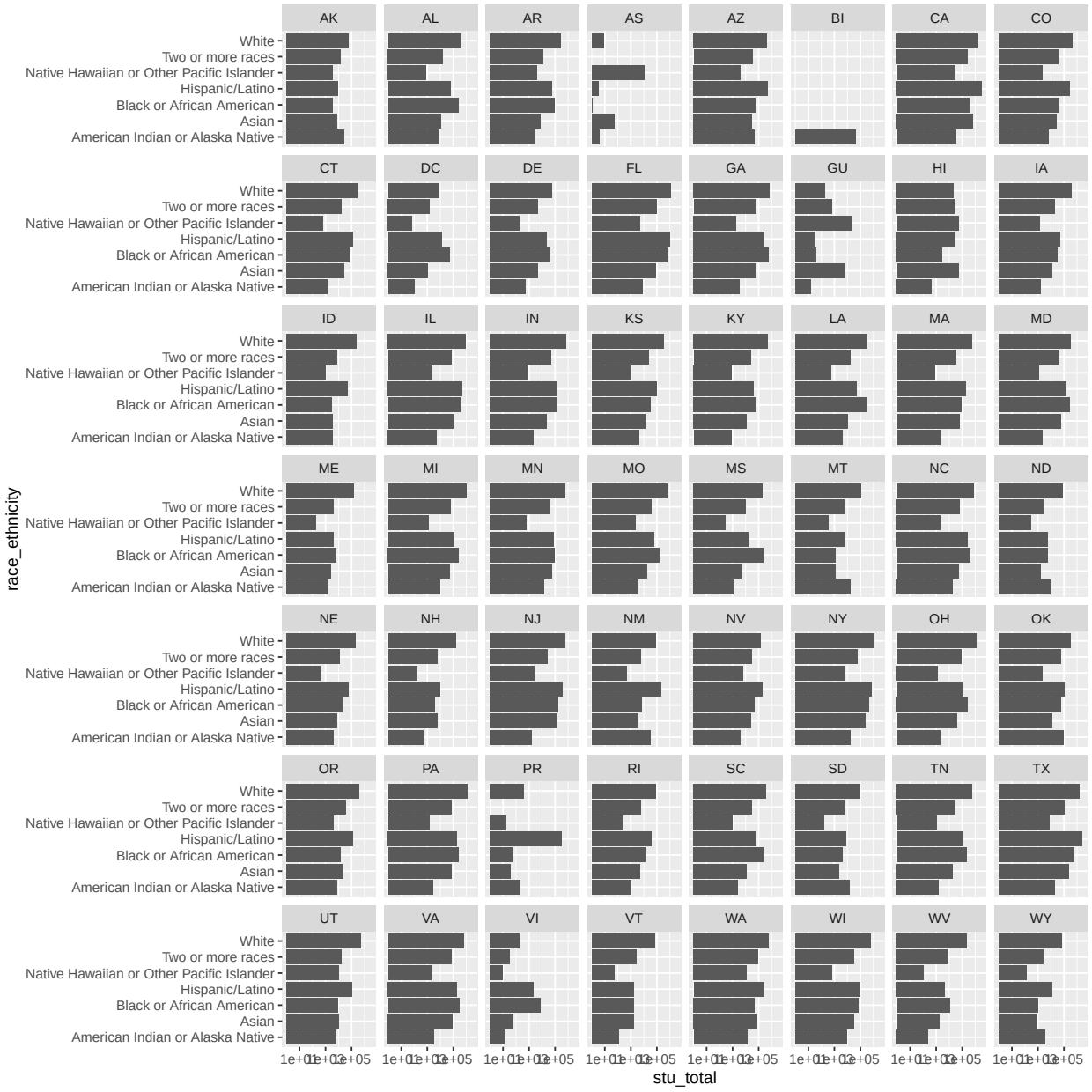
```
dm18_long %>%
    group_by(st, race_ethnicity) %>%
    summarise(
        stu_total = sum(student)
    ) %>%
    ggplot(aes(x = race_ethnicity, y = stu_total))+
    geom_col()+
    coord_flip()+
    facet_wrap(~st)
```



Plot 1: Faceted Bar Charts

```
#Some variation is seem, but most numbers are collapsed due to common x-axis.
#trying log transformation and percentage
```

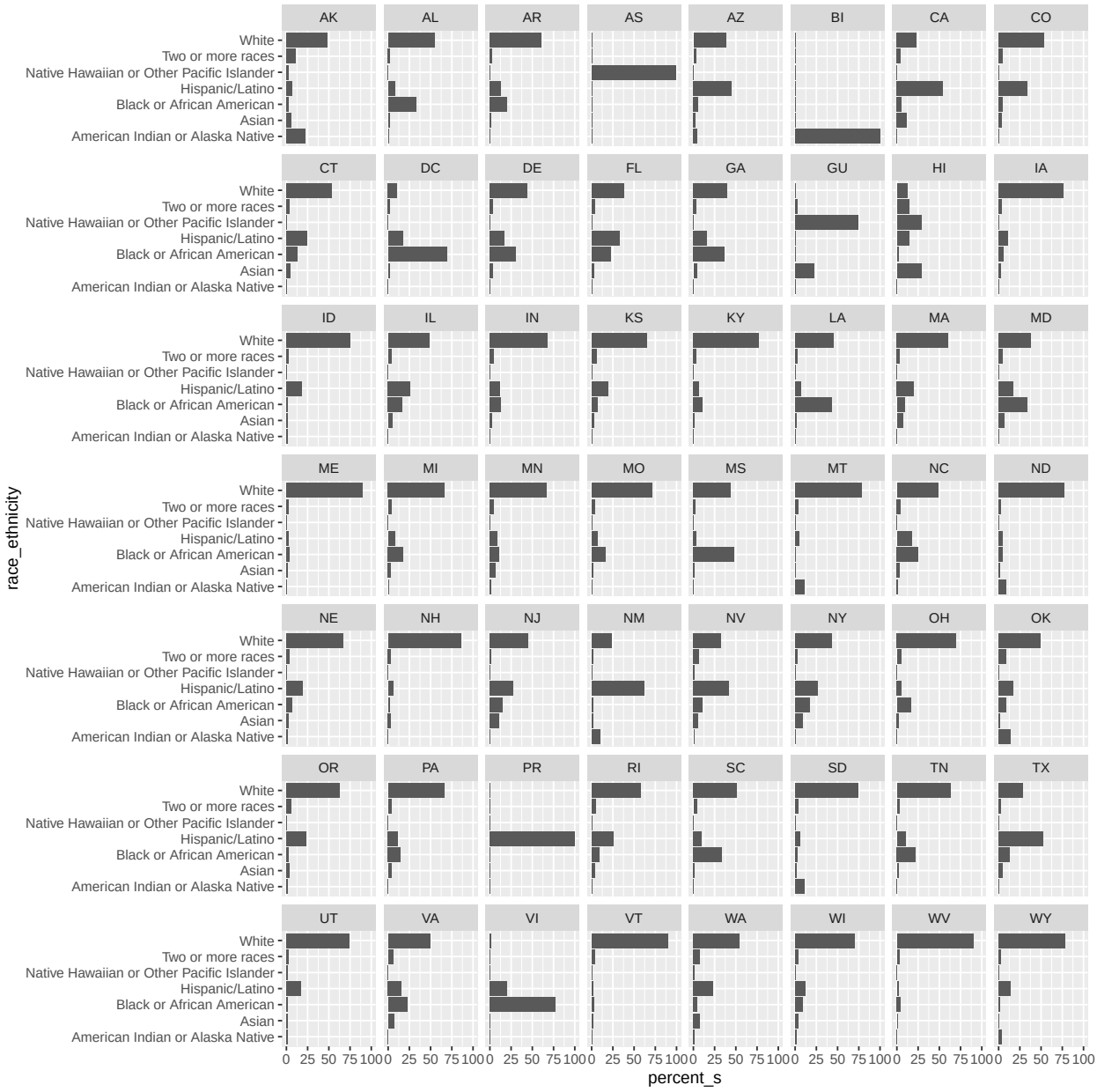
```
dm18_long %>%
    group_by(st, race_ethnicity) %>%
    summarise(
        stu_total = sum(student)
    ) %>%
    ggplot(aes(x = race_ethnicity, y = stu_total))+
    geom_col()+
    scale_y_log10()+
    coord_flip()+
    facet_wrap(~st)
```




```
#Some variation, but still hard to make much sense due to log scale
```

```
dm18_long %>%
    group_by(st, race_ethnicity) %>%
    summarise(
        stu_total = sum(student)
    ) %>%
    group_by(st) %>%
    mutate(
        total = sum(stu_total),
        percent_s = round((stu_total*100/total),3)
    ) %>%
    ggplot(aes(x = race_ethnicity, y = percent_s))+
```

```
    geom_col()+
    coord_flip()+
    facet_wrap(~st)
```




```
#this turned out well.
```

```
dm18_long %>%
    group_by(st, race_ethnicity) %>%
    summarise(
        stu_total = sum(student)
```

```
    ) %>%
    group_by(st) %>%
    mutate(
        total = sum(stu_total),
        percent_s = round((stu_total*100/total),3)
    ) %>%
    ggplot(aes(x = race_ethnicity,
               y = percent_s,
               fill = race_ethnicity))+
    geom_col(show.legend = FALSE)+
    coord_flip()+
    theme_bw() +
    facet_geo(~ st, grid = "us_state_grid2")
```

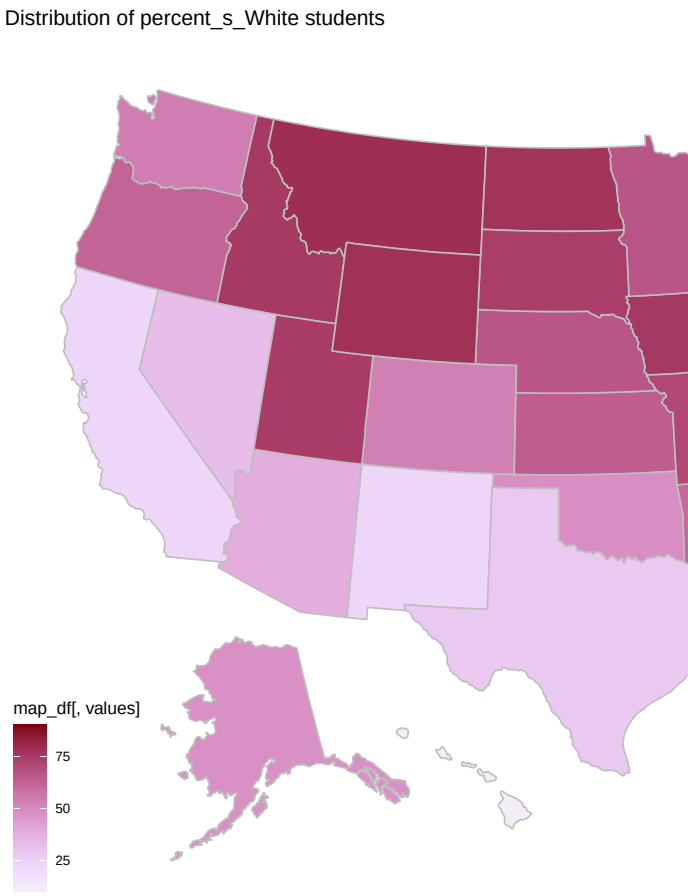
Plot 2: Bar charts with different layout

```
#Prep data for state (still can't figure out district mapping) in pivot_wider format

sm_2018 <- dm18_long %>%
```

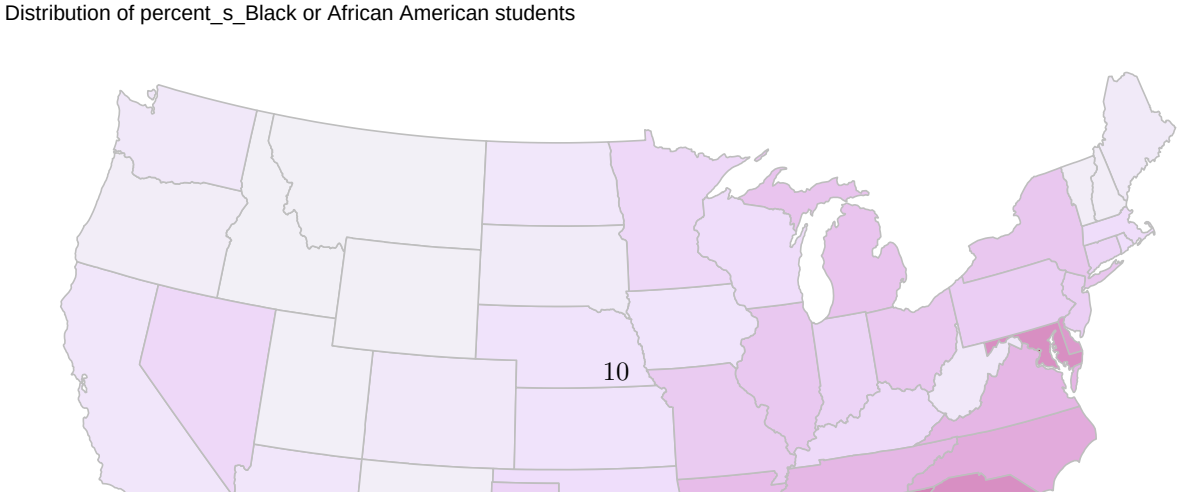
```
    group_by(st, statename, race_ethnicity) %>%
    summarise(
        stu_total = sum(student)
    ) %>%
    group_by(st) %>%
    mutate(
        total = sum(stu_total),
        percent_s = round((stu_total*100/total),3),
        state = tolower(statename)
    ) %>%
    pivot_wider(
        names_from = race_ethnicity,
        values_from = c(stu_total,percent_s)
    )

#check <- left_join(statepop, sm_2018, by = c("abbr" = "st")) %>%
 #   select(-pop_2015)
```

Plot 3: Maps

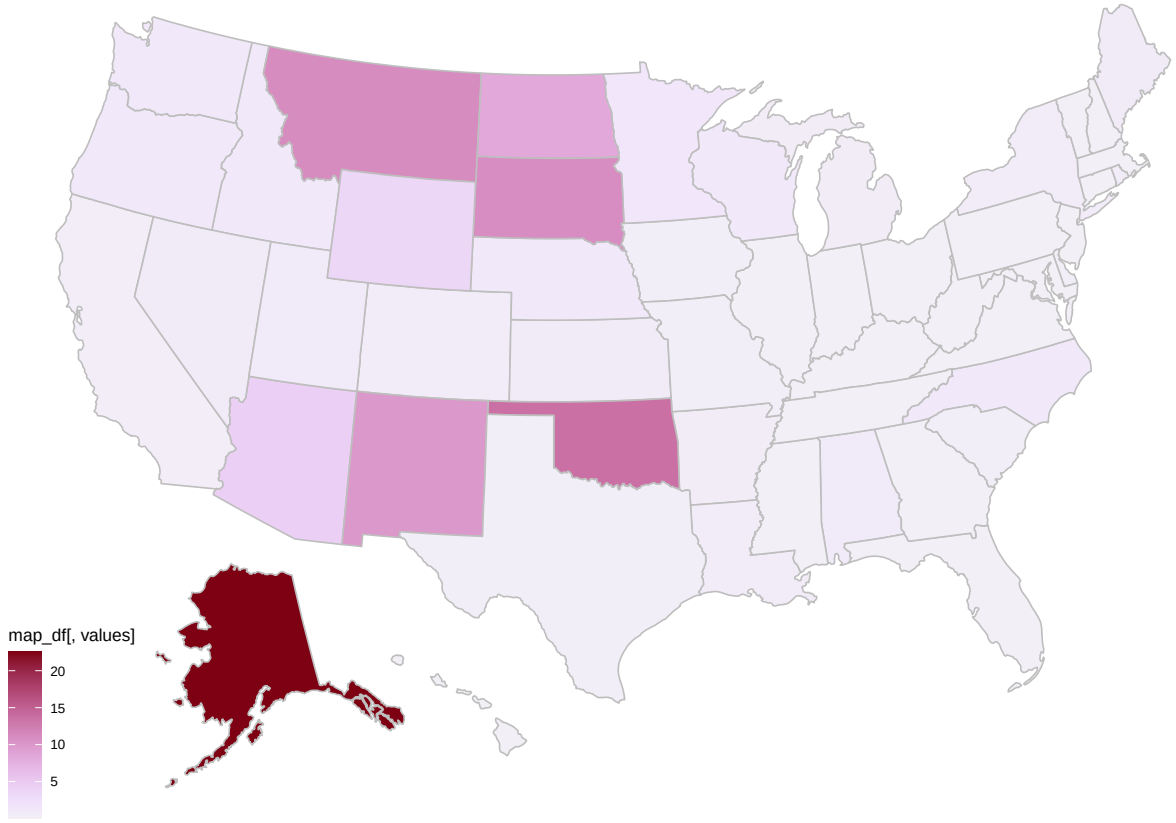
```
race <- c("percent_s_White","percent_s_Black or African American", "percent_s_American Indian or Alaska

plots <- vector("list", length(race))


for (i in seq_along(race)) {
plots[[i]] <- plot_usmap(data = sm_2018, values = race[i], color = "gray")+
    scale_fill_continuous_sequential(palette = "Red-Purple") +
    labs(
        title = paste0("Distribution of ", race[i]," students")
        )
print(plots[[i]])
}
```



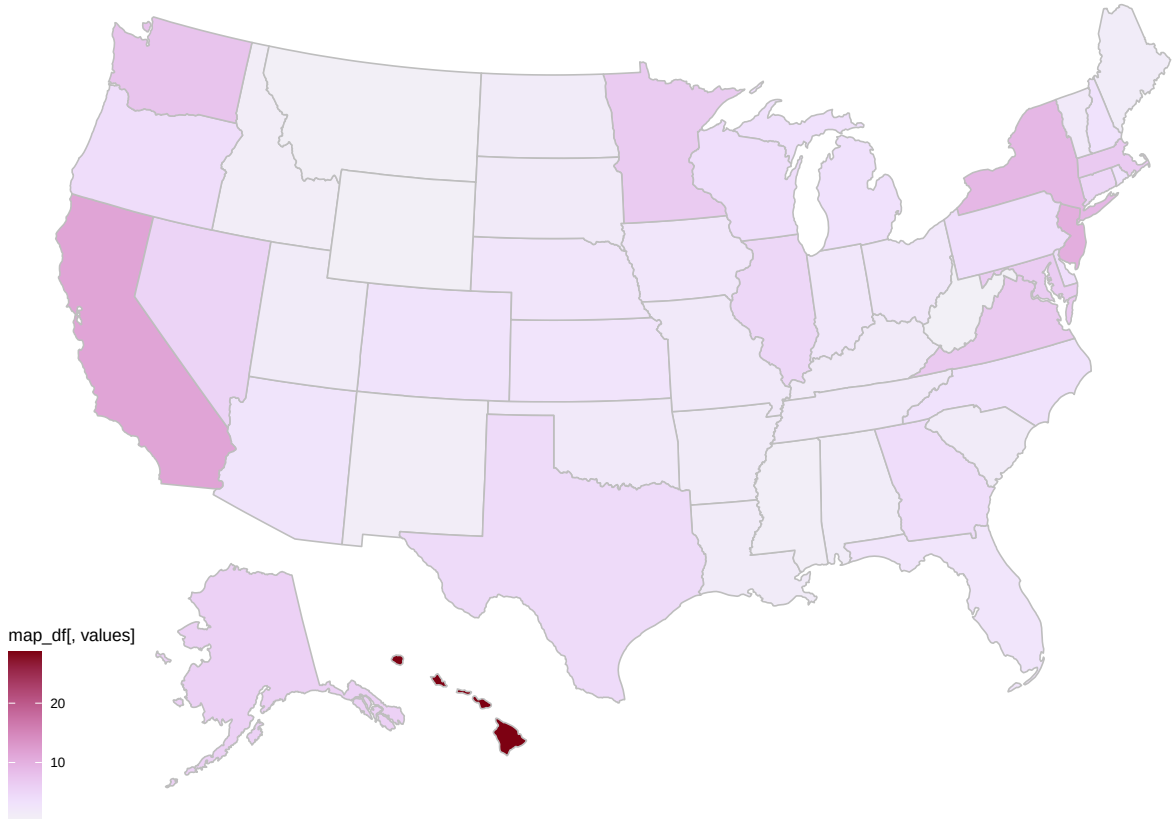
Plot 3: Geographic maps for student ethnic percentage



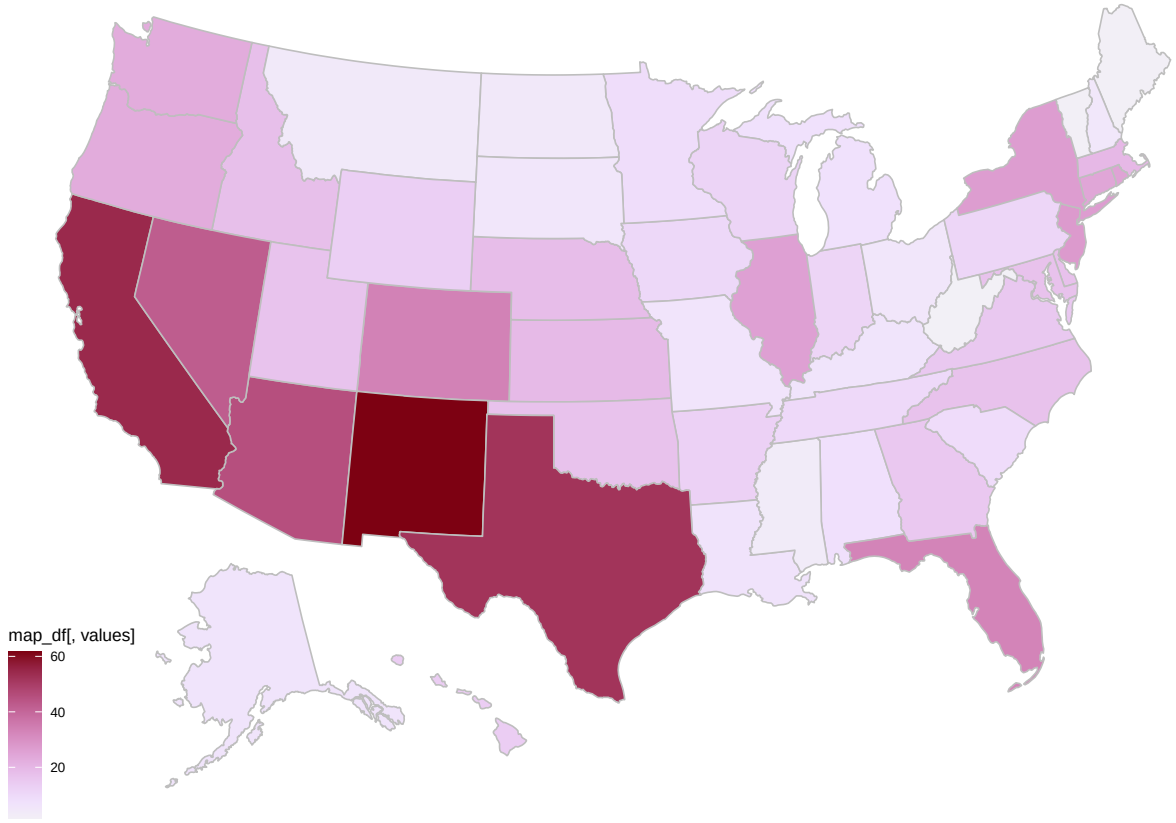
Distribution of percent_s_American Indian or Alaska Native students



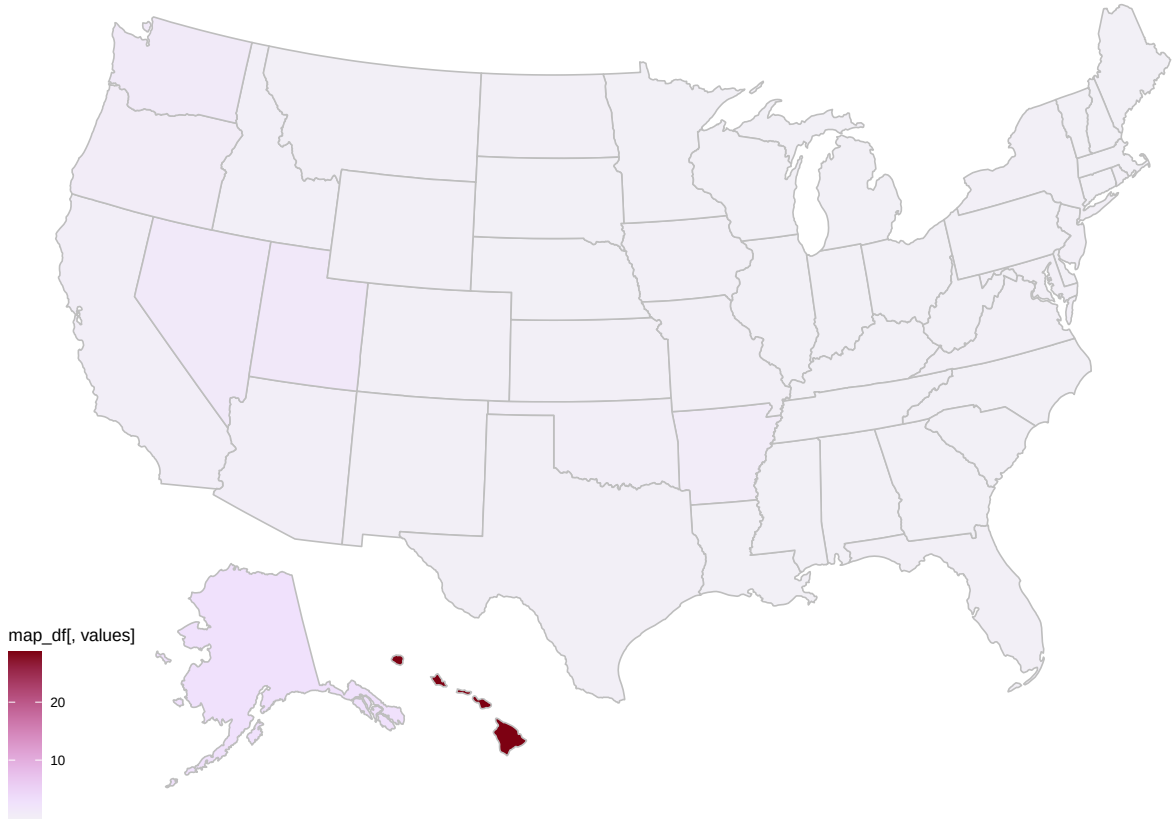
Distribution of percent_s_Asian students



Distribution of percent_s_Hispanic/Latino students

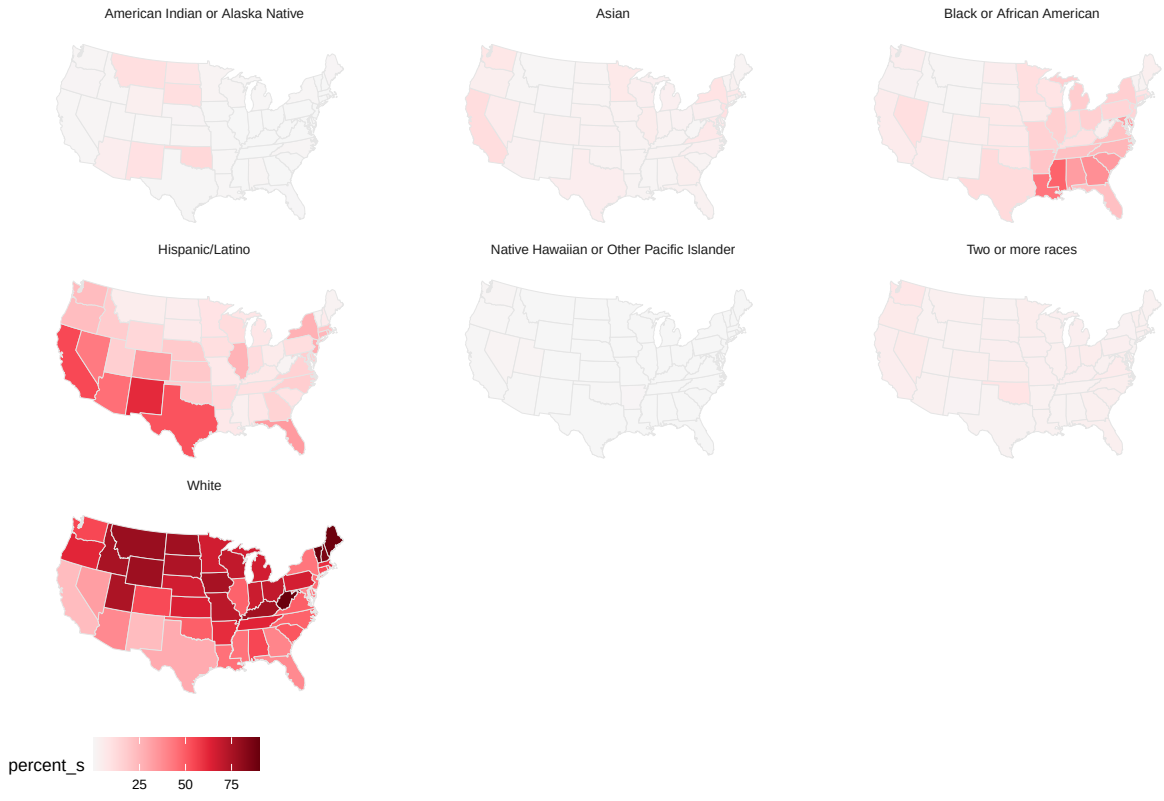


Distribution of percent_s_Native Hawaiian or Other Pacific Islander students

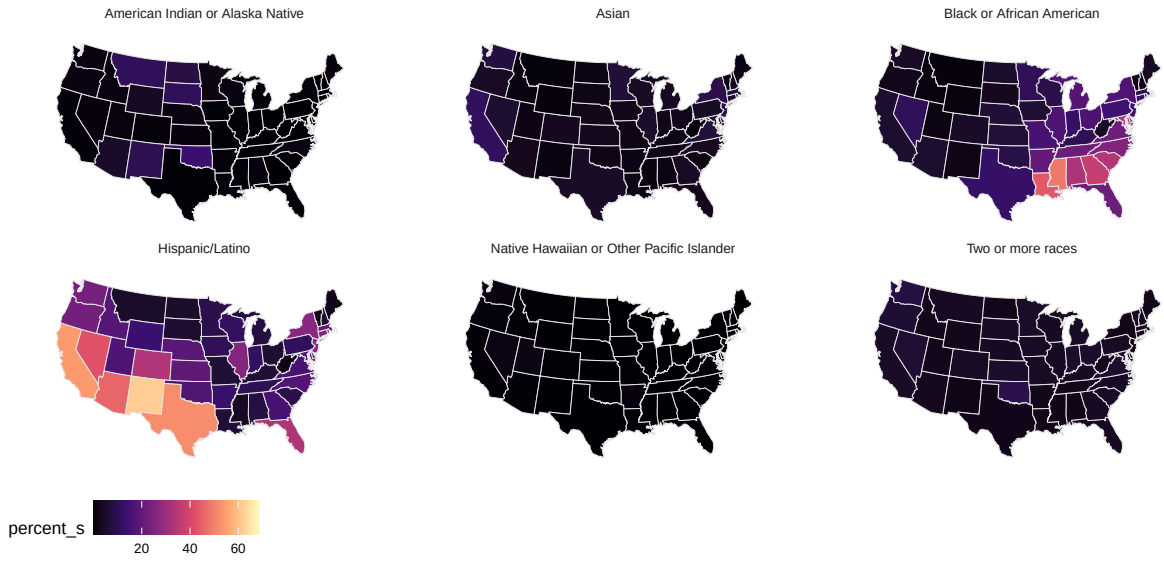



```
us_states <- map_data("state")

sm18_long <- dm18_long %>%
    group_by(st, statename, race_ethnicity) %>%
    summarise(
        stu_total = sum(student)
    ) %>%
    group_by(st) %>%
    mutate(
        total = sum(stu_total),
        percent_s = round((stu_total*100/total),3)
    ) %>%
    filter(!st %in% c("BI","AS","GU","PR","VI"))
```

```r
sm18_long$region <- tolower(sm18_long$statename)
stueth_map <- left_join(us_states, sm18_long)

ggplot(data = stueth_map,
             mapping = aes(x = long, y = lat,
                 group = group,
                 fill = percent_s))+
    geom_polygon(color = "gray90", size = 0.05) +
    coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
    scale_fill_continuous_sequential(palette = "Reds 3") +
    theme_map() + facet_wrap(~ race_ethnicity, ncol = 3) +
    theme(legend.position = "bottom",
          strip.background = element_blank())
```

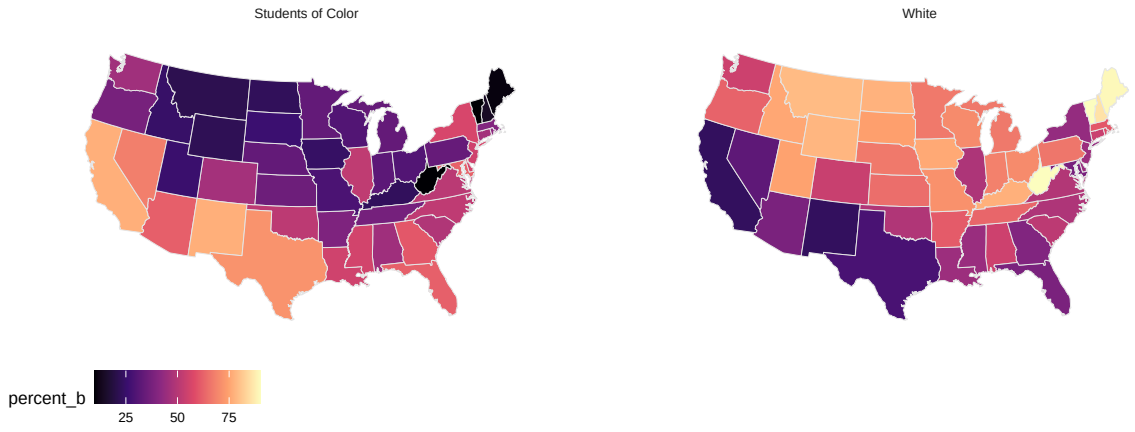


```
ggplot(data = subset(stueth_map, race_ethnicity != "White"),
             mapping = aes(x = long, y = lat,
                 group = group,
                 fill = percent_s))+
    geom_polygon(color = "gray90", size = 0.05) +
    coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
    scale_fill_viridis_c(option = "magma")+
    theme_map() + facet_wrap(~ race_ethnicity, ncol = 3) +
    theme(legend.position = "bottom",
          strip.background = element_blank())
```

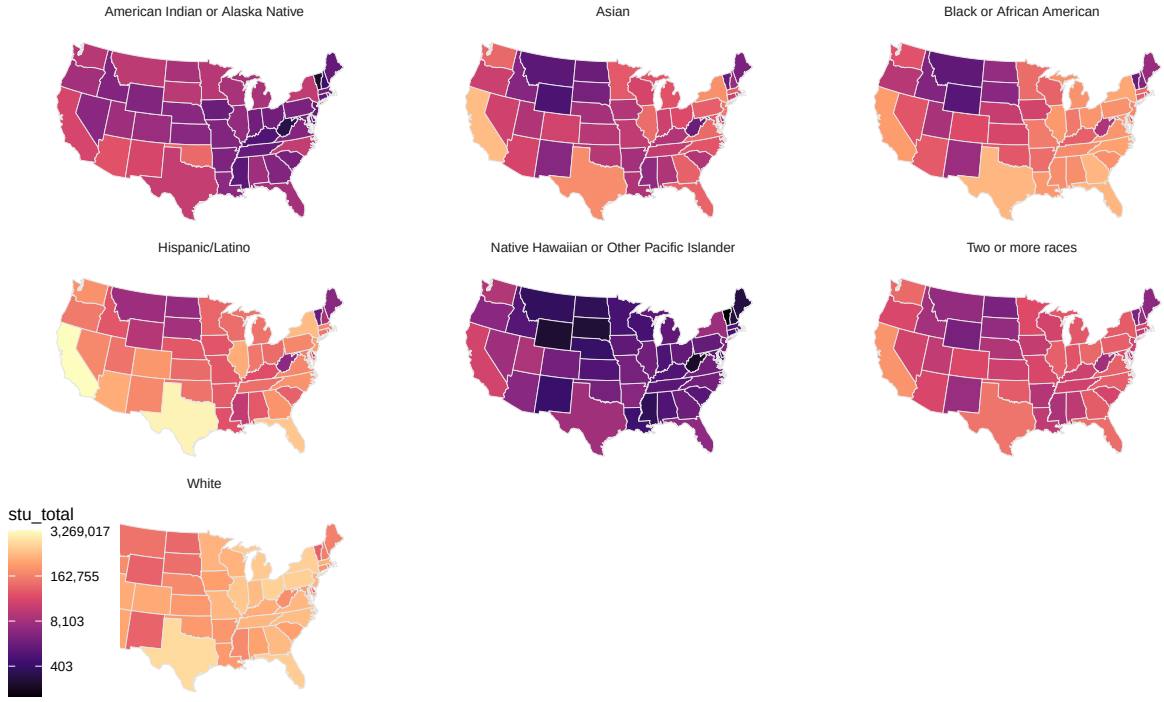



```
sm18_long %>%
    mutate(
```

```
    rbinary = ifelse(race_ethnicity == "White", "White","Students of Color")
) %>%
group_by(region, rbinary) %>%
summarise(
    percent_b = sum(percent_s)
) %>%
right_join(us_states) %>%
ggplot(mapping = aes(x = long, y = lat,
             group = group,
             fill = percent_b))+
geom_polygon(color = "gray90", size = 0.05) +
coord_map(projection = "albers", lat0 = 39, lat1 = 45)+
scale_fill_viridis_c(option = "magma")+
theme_map() + facet_wrap(~ rbinary, ncol = 2) +
theme(legend.position = "bottom",
      strip.background = element_blank())
```




```
#also trying logarithmic scale for raw counts of student ethnicity
ggplot(data = stueth_map,
             mapping = aes(x = long, y = lat,
                 group = group,
                 fill = stu_total))+
    geom_polygon(color = "gray90", size = 0.05) +
    coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
    scale_fill_viridis_c(option = "magma",
                         trans = "log",
                         labels = scales::comma)+
    theme_map() + facet_wrap(~ race_ethnicity, ncol = 3) +
    theme(strip.background = element_blank(),
          legend.direction = "vertical")
```



Research Question 2

2a. How does the the proficiently level in language and math vary across state for High School Students? How does it differ by students characteristics such as race/ethnicity, English Learner status, Student with Disability, Low Income students?

```
rla_sc <- get_data("EDFacts_rla_achievement_sch_2010_2019")
#rla_sc <- import(here("Data", "rla_sc.csv"))  #uncomment this read from your local.
fis08 <- import(here("Data","fis08.csv"))
math_sc <- import(here("Data", "math_sc.csv"))
enroll <- import(here("Data", "dm_total.rda"))
#rla_sc <- import(here("Data", "rla_sc.csv")) file is too big, I can't push it to github
```

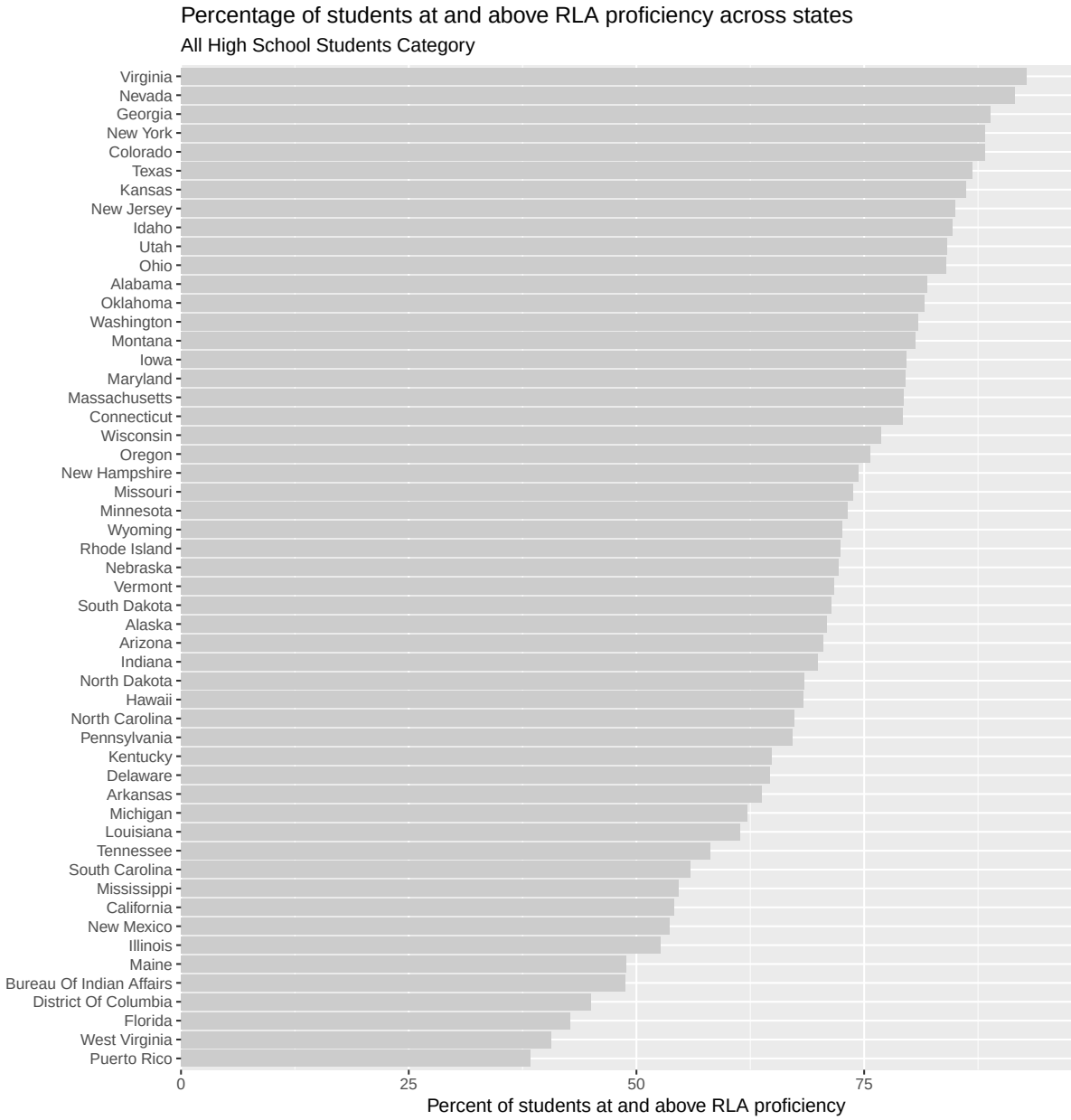
```
math_sc <- math_sc %>%
  select(LEAID, STNAM, NCESSCH, ALL_MTHHSPCTPROF, MAM_MTHHSPCTPROF, MAS_MTHHSPCTPROF, MBL_MTHHSPCTPROF,
  clean_names() %>%
  pivot_longer(cols= ends_with("prof"),
               names_to = "identity",
               values_to = "math_pctabove",
               names_pattern = "(.*)_mthhspctprof")

math_sc$math_pctabove <- sub(".*-(.*)", "\\1", math_sc$math_pctabove)
math_sc$math_pctabove <- as.numeric(math_sc$math_pctabove)
math_sc$leaid <- as.character(math_sc$leaid)
math_sc$ncessch <- as.character(math_sc$ncessch)


rla_sc <- rla_sc %>%
   select(LEAID,STNAM, NCESSCH, ALL_RLAHSPCTPROF, MAM_RLAHSPCTPROF, MAS_RLAHSPCTPROF, MBL_RLAHSPCTPROF,
  clean_names() %>%
  pivot_longer(cols= ends_with("prof"),
               names_to = "identity",
               values_to = "rla_pctabove",
               names_pattern = "(.*)_rlahspctprof")
rla_sc$rla_pctabove <- sub(".*-(.*)", "\\1", rla_sc$rla_pctabove)
rla_sc$rla_pctabove <- as.numeric(rla_sc$rla_pctabove)


fis08 <- fis08 %>%
  select(LEAID, "textbook"= V93, TOTALEXP, "instruction"= TCURINST , "supservice" = TCURSSVC ) %>%
  clean_names()
```

```
all <- left_join(rla_sc, math_sc) %>%
  left_join(fis08)

enroll <- enroll %>%
  select(LEAID, LEA_NAME, STUDENT_COUNT)%>%
  clean_names()



all<- left_join(all,enroll) %>%
  mutate(perstudbook = (textbook/student_count),
         stnam = stringr::str_to_title(stnam)) %>%
  filter_all(all_vars(!is.infinite(.))) %>%
  filter(!stnam == "Stnam")
```

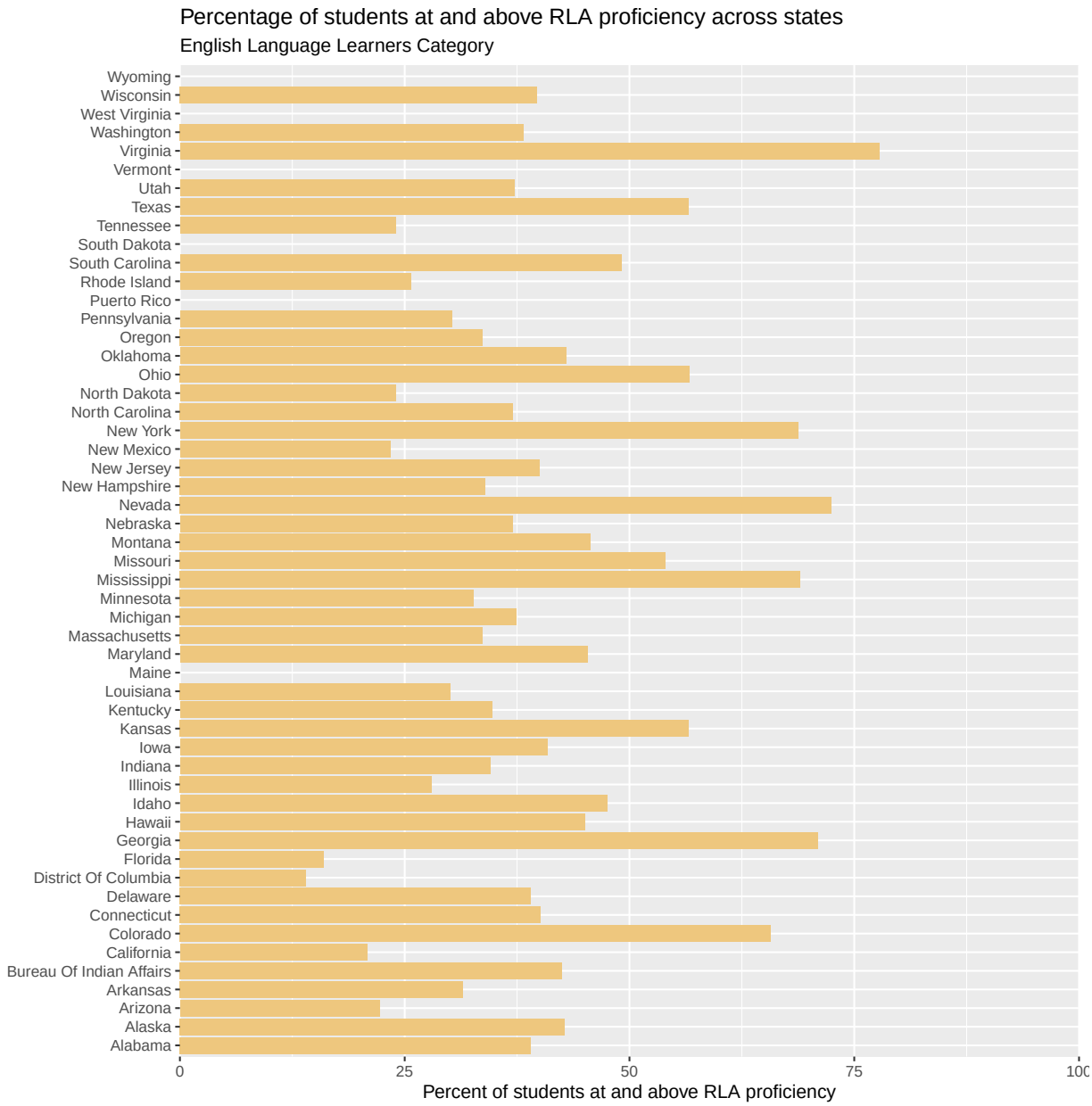
```
#Planning to have interactive plots with all student as gray background and other characteristics with
#all students
all %>%
  filter(identity == "all") %>%
  group_by(stnam) %>%
  summarise(meanmath = mean(math_pctabove, na.rm = TRUE),
```

```
          meanrla = mean(rla_pctabove, na.rm=TRUE)) %>%
 ggplot()+
geom_col(aes(meanrla, fct_reorder(stnam, meanrla)), fill = "grey80") +
 scale_x_continuous(expand = c(0,0),
                    limits = c(0,100) ) +
labs(title = "Percentage of students at and above RLA proficiency across states",
    subtitle = "All High School Students Category",
      y = "",
    x= "Percent of students at and above RLA proficiency")
```

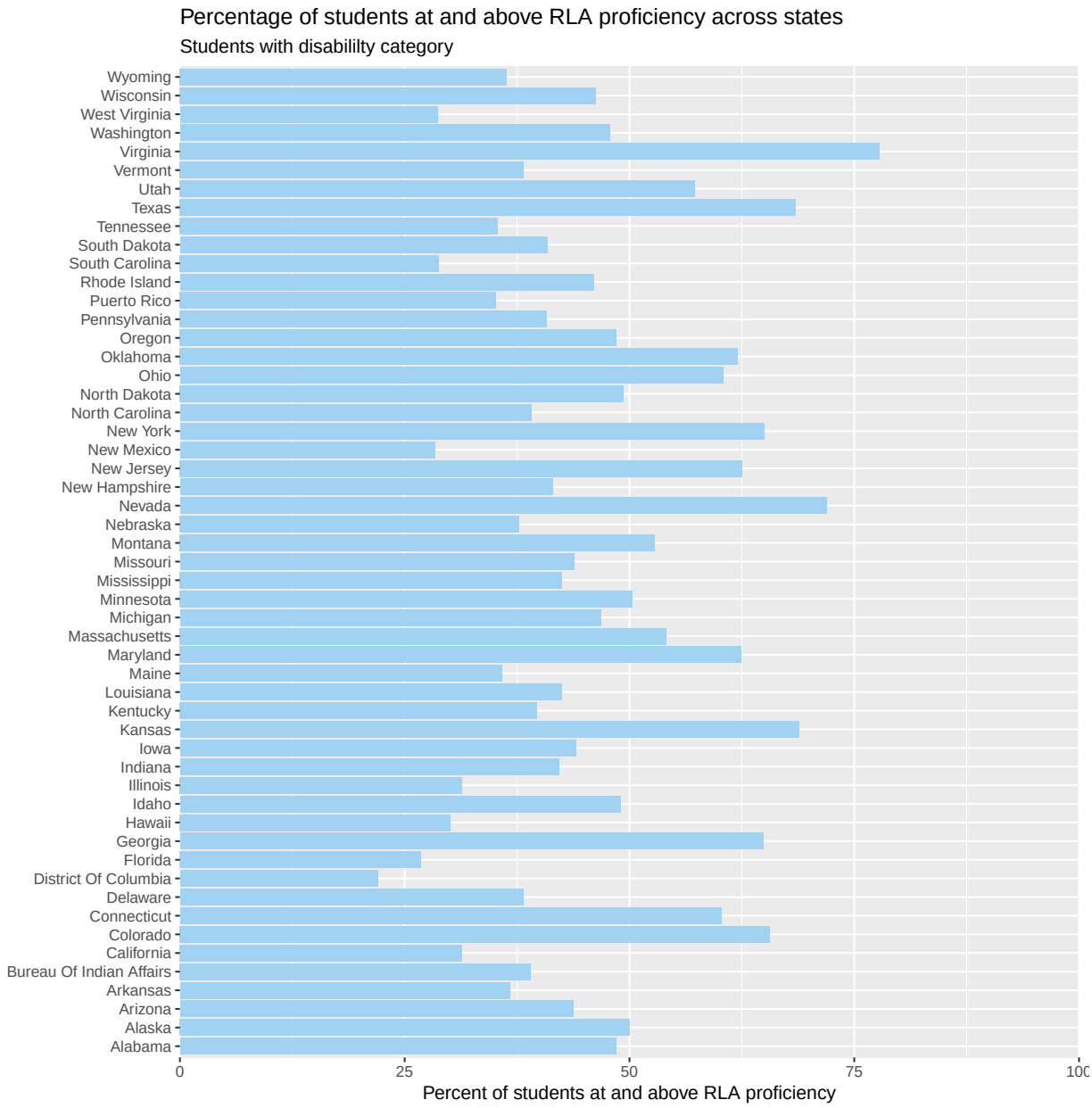
Plots to compare students proficiency level across states and how it differs based on students' characteristics.



```r
#proficiency level for english language learner
all %>%
  filter(identity == "lep") %>%
  group_by(stnam) %>%
  summarise(meanmath = mean(math_pctabove, na.rm = TRUE),
            meanrla = mean(rla_pctabove, na.rm=TRUE)) %>%
  ggplot()+
  geom_col(aes(meanrla, stnam), fill = "#EEC77E")+
  scale_x_continuous(expand = c(0,0),
                     limits = c(0,100))+
  labs(title = "Percentage of students at and above RLA proficiency across states",
       subtitle = "English Language Learners Category",
         y = "",
       x= "Percent of students at and above RLA proficiency")
```

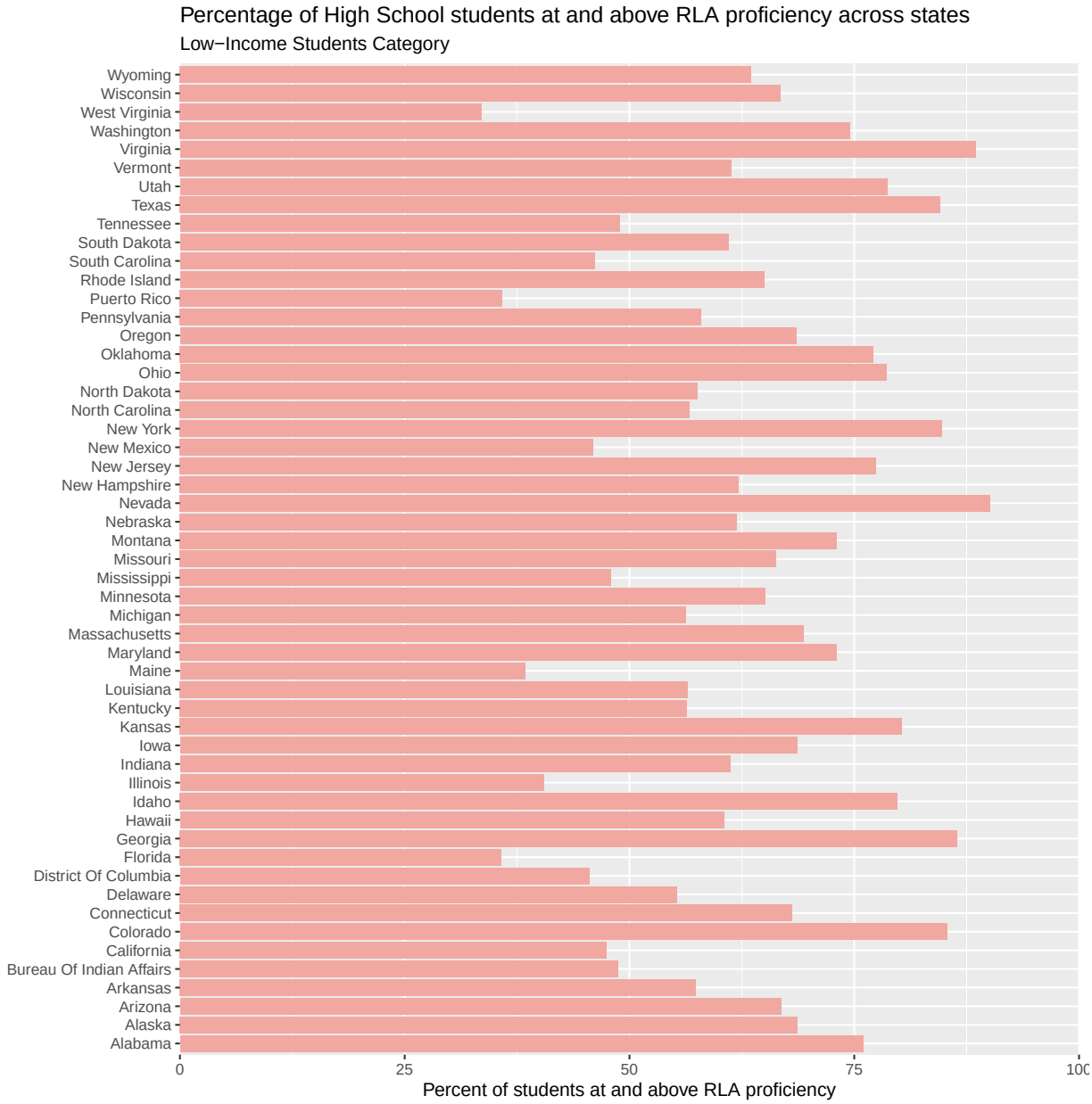



```
#proficiency level for children with disability
all %>%
  filter(identity == "cwd") %>%
  group_by(stnam) %>%
  summarise(meanmath = mean(math_pctabove, na.rm = TRUE),
            meanrla = mean(rla_pctabove, na.rm=TRUE)) %>%
  ggplot()+
  geom_col(aes(meanrla, stnam), fill = "#A1D2F1")+
  scale_x_continuous(expand = c(0,0),
                     limits = c(0,100))+
    labs(title = "Percentage of students at and above RLA proficiency across states",
       subtitle = "Students with disabililty category",
         y = "",
       x= "Percent of students at and above RLA proficiency")
```




```
#proficiency level for Economically disadvantaged students
all %>%
  filter(identity == "ecd") %>%
  group_by(stnam) %>%
  summarise(meanmath = mean(math_pctabove, na.rm = TRUE),
            meanrla = mean(rla_pctabove, na.rm=TRUE)) %>%
  ggplot()+
  geom_col(aes(meanrla, stnam), fill = "#F1A8A1")+
  scale_x_continuous(expand = c(0,0),
                     limits = c(0,100)) +
  labs(title = "Percentage of High School students at and above RLA proficiency across states",
       subtitle = "Low-Income Students Category",
```

```
    y = "",
  x= "Percent of students at and above RLA proficiency")
```

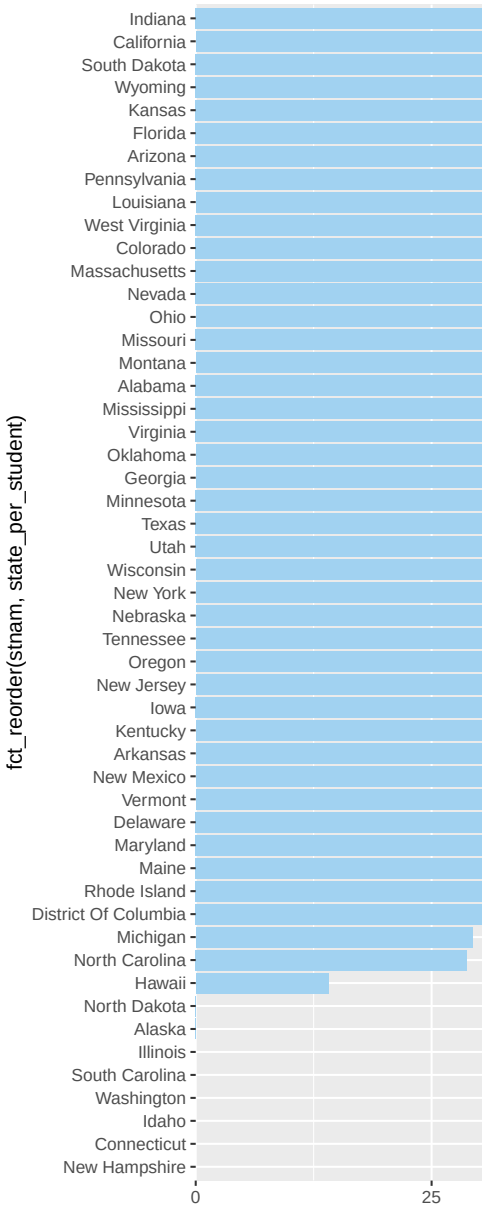


2b. What is the relationship between district spending on textbook and students proficiency level?

```
#State level data
state_textbook <- all %>%
  select(stnam, textbook, student_count) %>%
  distinct(stnam, textbook, student_count) %>%
  drop_na() %>%
```

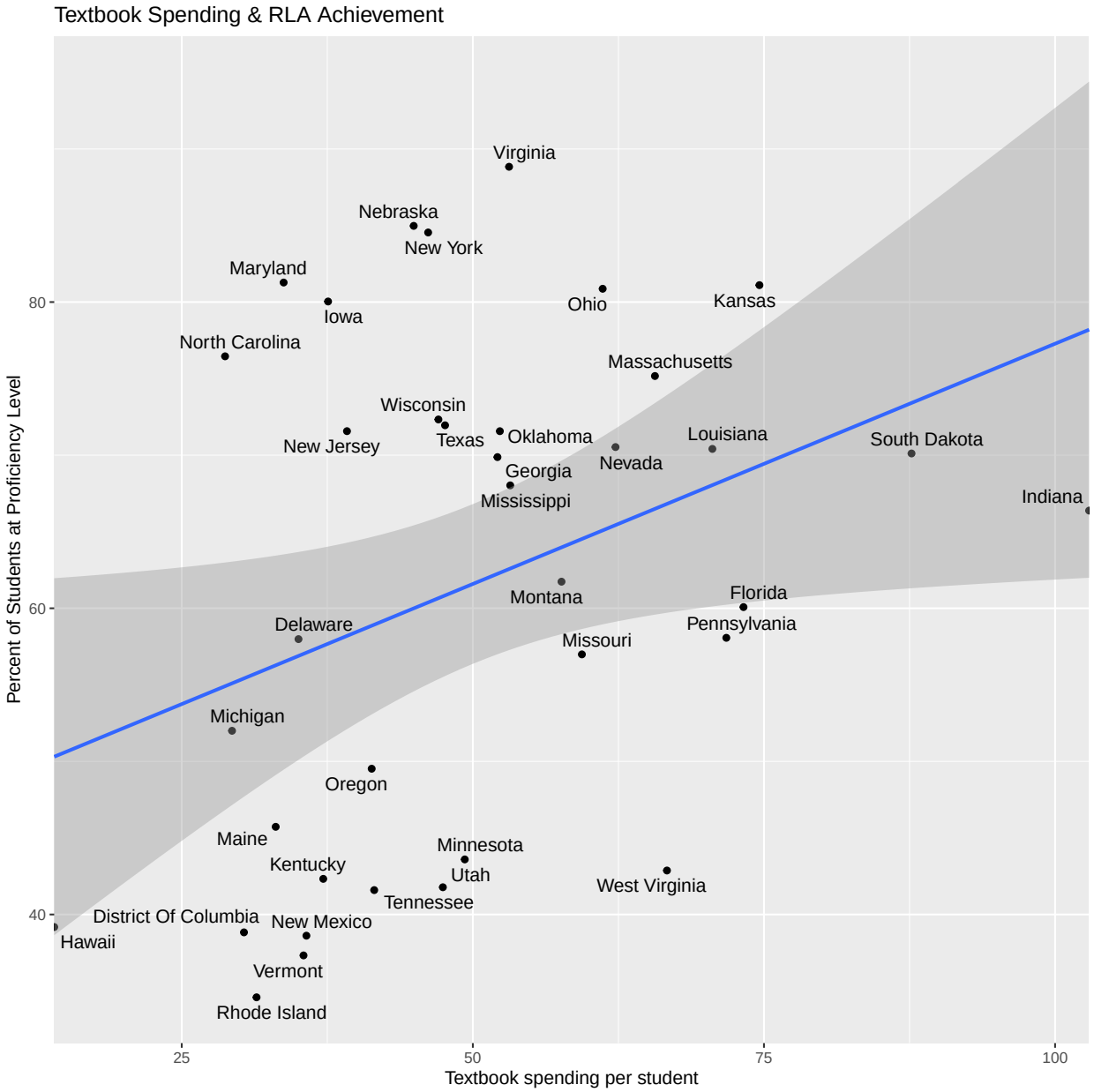
```
  group_by(stnam) %>%
  summarise(state_per_student = sum(textbook)/sum(student_count))


state_pct<- all %>%
  filter(identity == "all") %>%
  group_by(stnam) %>%
  summarise(meanmath = mean(math_pctabove, na.rm = TRUE),
            meanrla = mean(rla_pctabove, na.rm=TRUE))

state_joined<-left_join(state_textbook, state_pct)
```

```
state_textbook  %>%
  ggplot()+
  geom_col(aes(state_per_student, fct_reorder(stnam, state_per_student)), fill = "#A1D2F1")+
  scale_x_continuous(expand = c(0,0),
                     limits = c(0,110)) +
  labs( x = "Textbook spending per student")
```

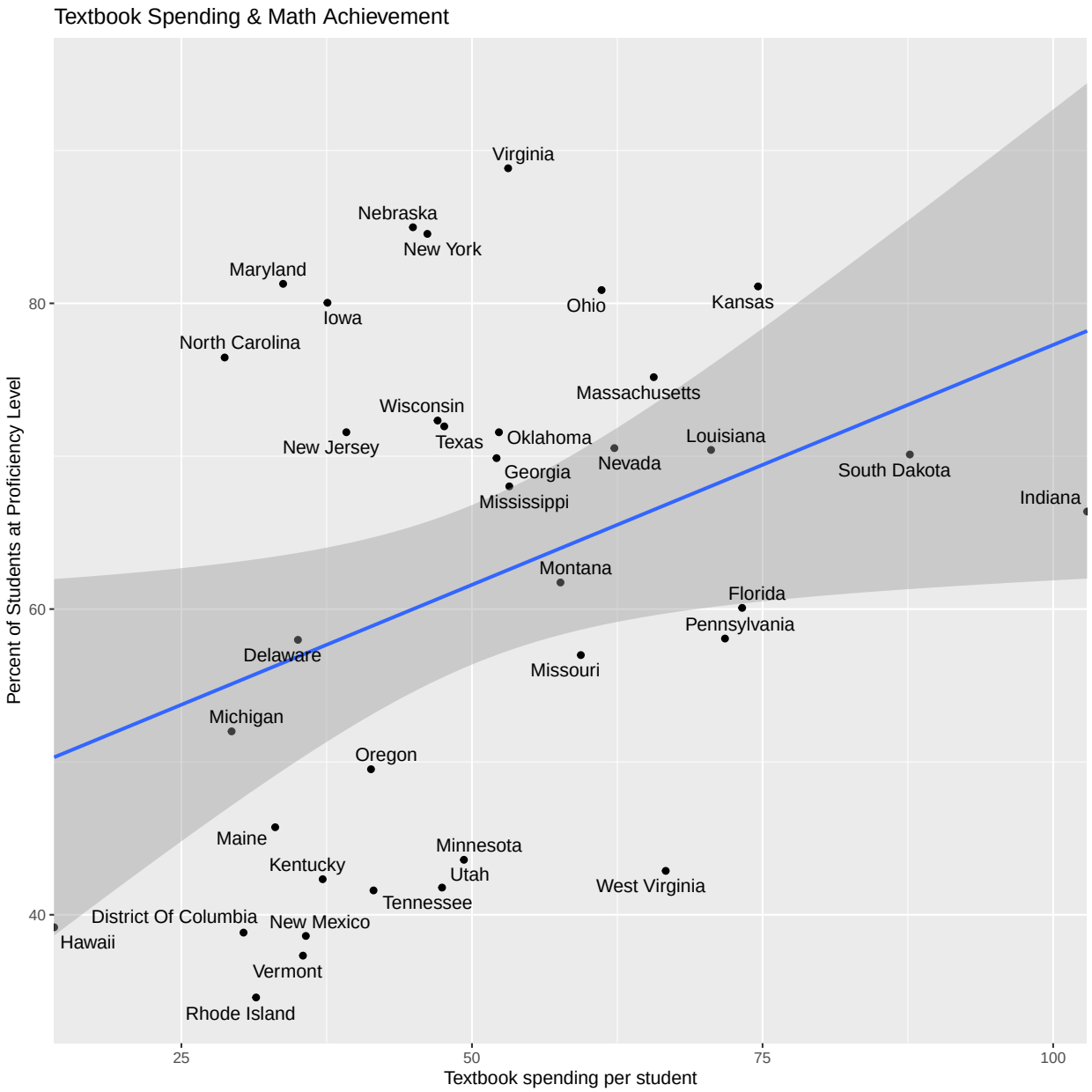


Relationship between Textbook Spending & RLA / Math Achievement

```
state_joined%>%
  filter(!state_per_student <0,
         !stnam == "North Dakota",
         !stnam == "Alaska") %>% #drop state with missing values
  ggplot(aes(state_per_student, meanmath))+
  geom_point() +
  geom_smooth(method = "lm")+
  geom_text_repel(aes(label = stnam))+
  scale_x_continuous(expand = c(0,0)) +
  labs(title= "Textbook Spending & RLA Achievement",
      x = "Textbook spending per student",
       y = "Percent of Students at Proficiency Level")
```



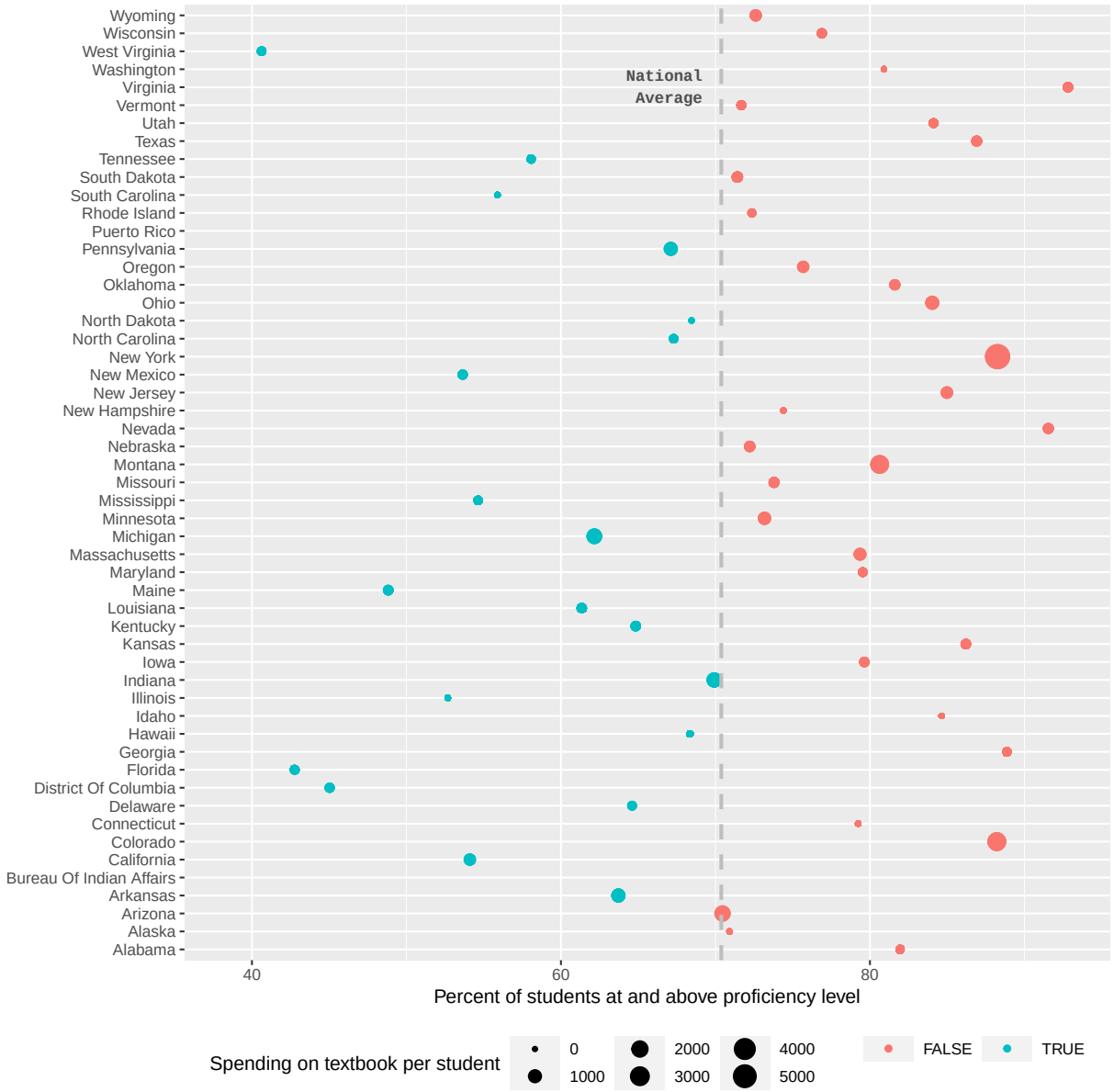

```
state_joined%>%
  filter(!state_per_student <0,
         !stnam == "North Dakota",
         !stnam == "Alaska") %>% #drop state with missing values
  ggplot(aes(state_per_student, meanmath))+
  geom_point() +
  geom_smooth(method = "lm")+
  geom_text_repel(aes(label = stnam))+
  scale_x_continuous(expand = c(0,0)) +
    labs(title= "Textbook Spending & Math Achievement",
      x = "Textbook spending per student",
       y = "Percent of Students at Proficiency Level")
```




```
cek <- all %>%
  filter(identity == "all") %>%
  group_by(stnam) %>%
  mutate(meanmath = mean(math_pctabove, na.rm = TRUE),
            meanrla = mean(rla_pctabove, na.rm=TRUE),
          low = ifelse(meanrla<=70.38, TRUE, FALSE)) %>%
  ggplot(aes(meanrla, stnam)) +
  geom_point(aes(size = perstudbook, color = low))+
  geom_vline(xintercept=70.38,
             linetype = "dashed",
             color = "gray",
```

```
              size = 1) +
   annotate("text",x=67 ,
            y = "Virginia" ,label = "National \nAverage",color = "gray30",size = 3,
            family="Courier", line = "gray", fontface="bold") +
  labs(title = "Relationship between Textbook Spending per Student and Language proficiency",
       x= "Percent of students at and above proficiency level",
       y="",
       size= "Spending on textbook per student",
       color="",
       legend= "") +
   theme(legend.position = "bottom",
         legend.direction = "horizontal",
         legend.key.size = unit(1, 'cm'),
         legend.key.height = unit(.5,"cm"),
         plot.title = element_text(hjust = 0),
         plot.title.position = "plot")

cek
```

Relationship between Textbook Spending per Student and Language proficiency Across the

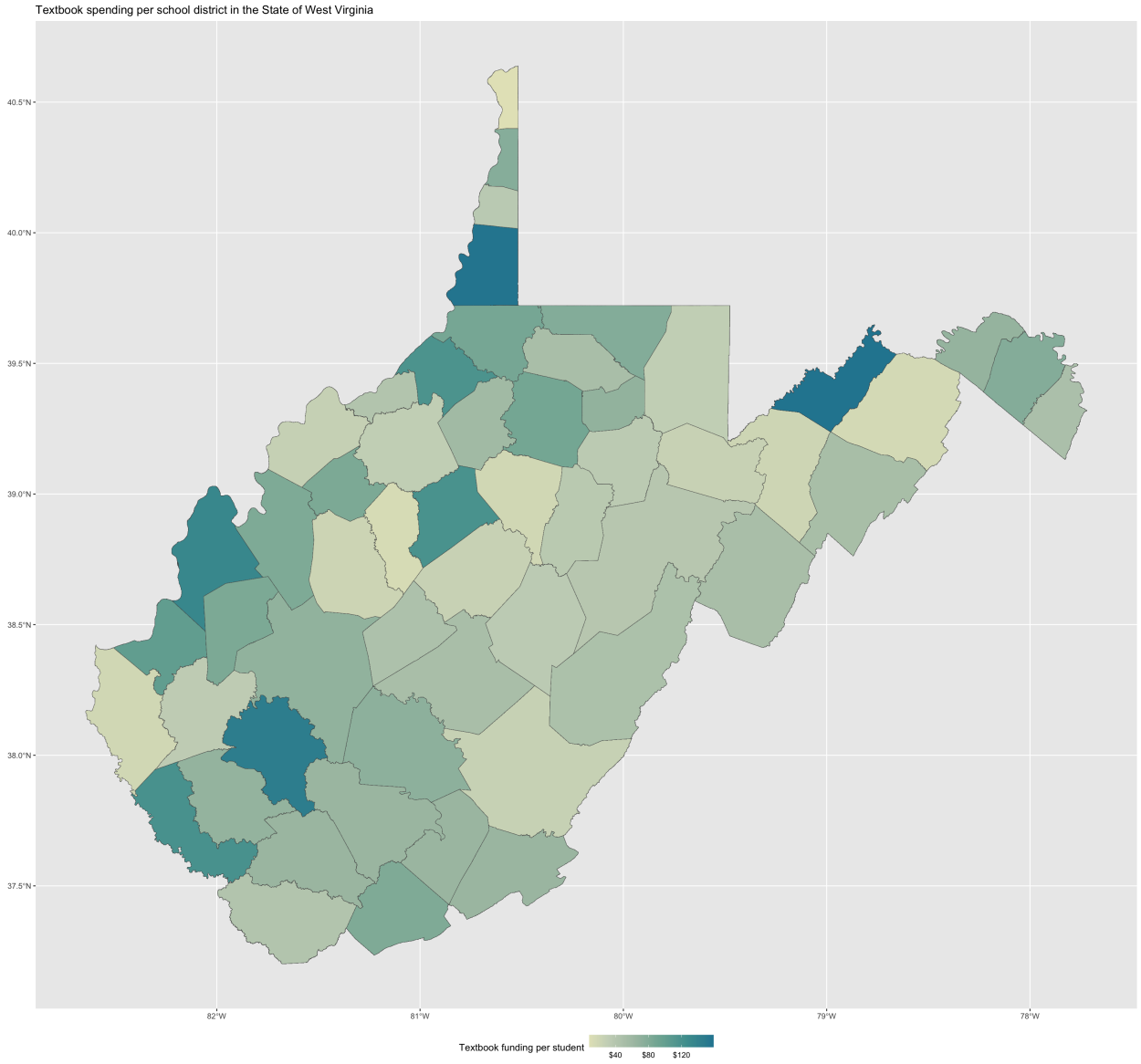


States

Variation in District Textbook Spending for each state

Example for West Virginia

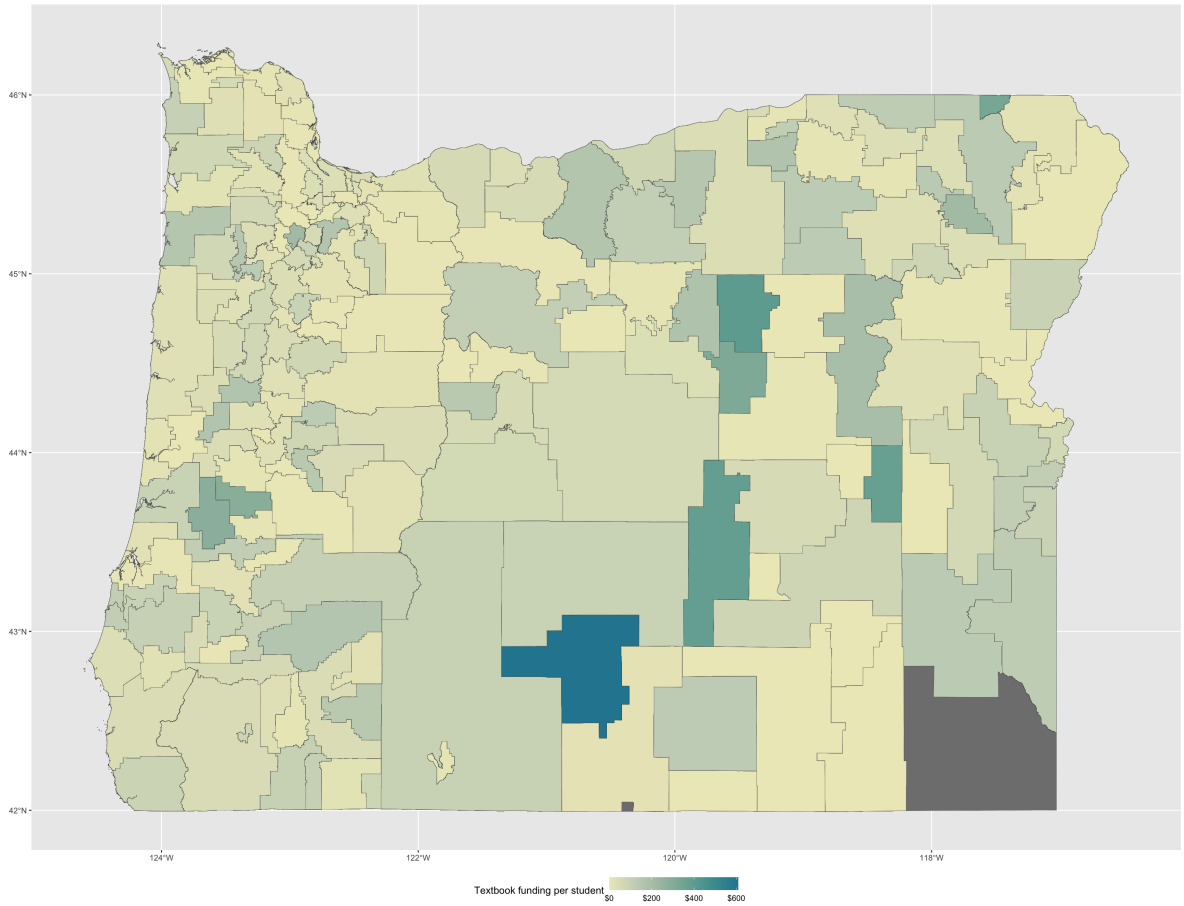
- Attaching the example plot as an image here because it requires specific setting to run.



{width = 8}

Visualizing textbook spending for school district in Oregon:

- Attaching the example plot as an image here because it requires specific setting to run.


46°N
45°N
44°N
43°N
42°N
124°W
122°W
120°W
118°W
Textbook funding per student
$0
$200
$400
$600

{width = 8}