HW#2

Lucía Carrera

9/24/2021

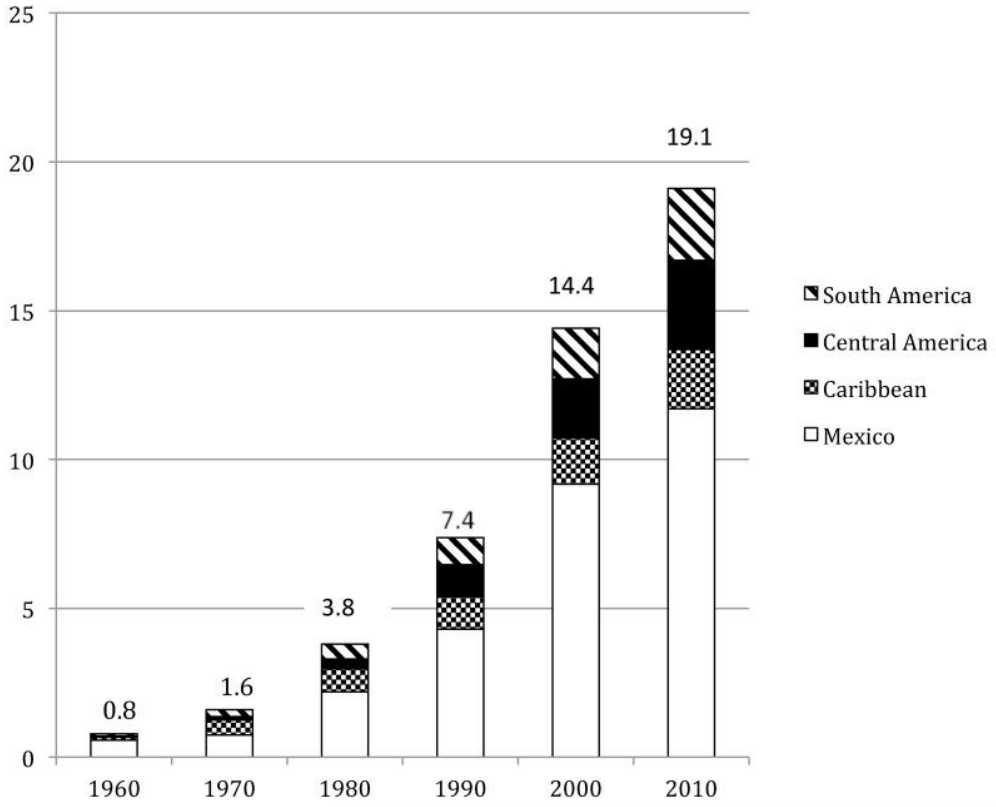
Set-up

```
knitr::opts_chunk$set(echo = TRUE)

#  Load the packages:  ggplot2, dplyr and babynames
#  Include loading code here:
library(ggplot2)
library(dplyr)
library(babynames)
```

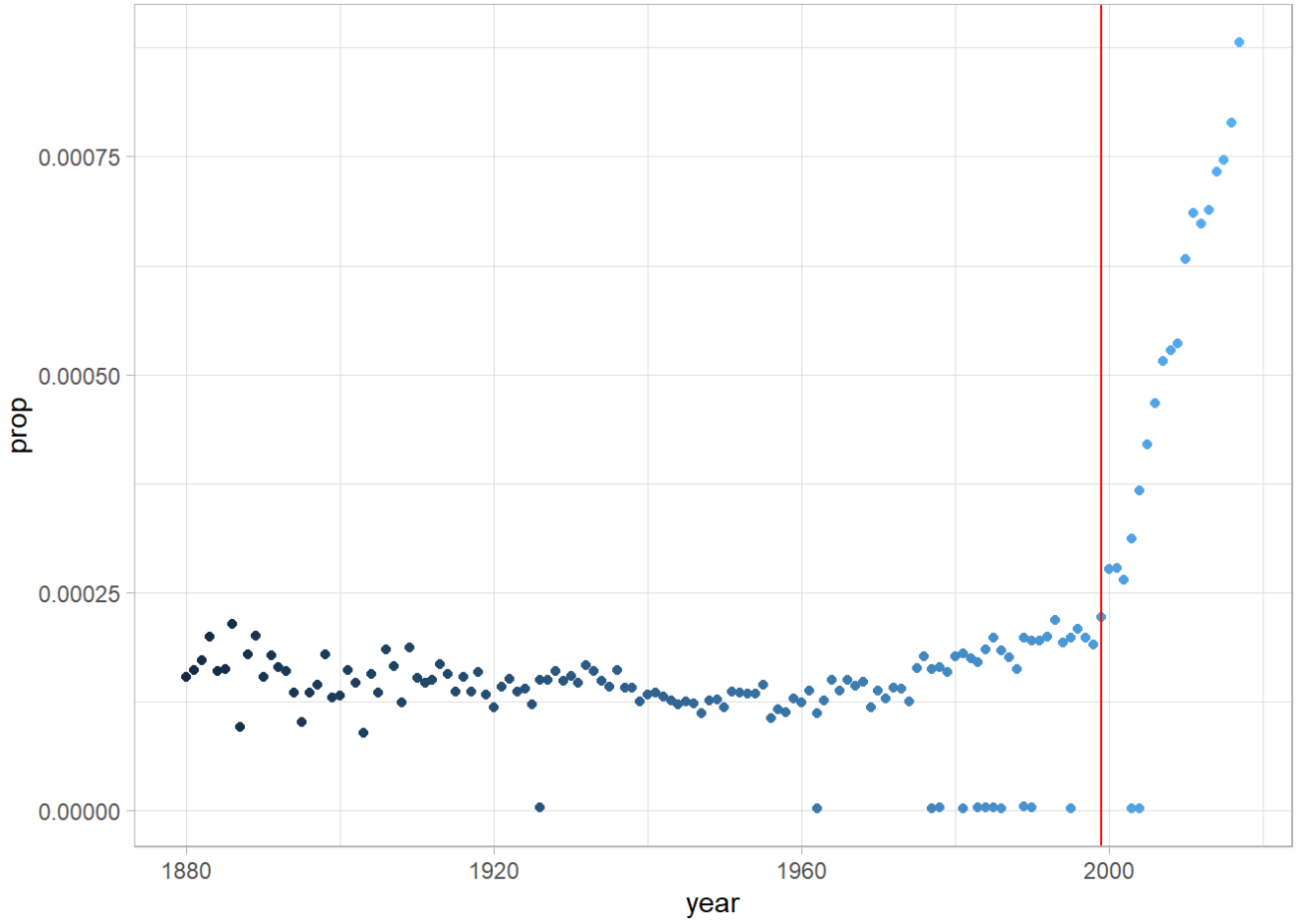
Question 1

The graph distinctly rises from the 2000 onwards. By using the help("babynames") function we can see that the data was provided by the Social Security Administration (SSA), an organization in the USA. I did this to narrow down the reason of to why it gained popularity (In Spain the reason to its popularity could be different to one in the US).

Given that Lucia is traditionally a Hispanic name (also Italian), I decided to search in the National Center for Biotechnology Information's website to see if immigration from Latin America had risen over the last decades. As seen in this manuscript (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4638184/), immigration from Latin American countries quintupled from the 1980s to the 2010s.



```
#  a.  Using the babynames data frame, create a new data frame with only your name* in it.  (* You may choose a different name, if you prefer, but not Sheila!)
Lucia <- babynames %>%
  filter(name == "Lucia")


#  b. Then graph the proportion of babies with your name (prop) by year using ggplot.  Use geom_line, and include a vertical line at the year of your birth, by adding this layer, putting in the year of your birth where it says 0000:     geom_vline(xintercept = 0000)
ggplot(data = Lucia,
       mapping = aes( x = year, y = prop, color = year )) +
  geom_point(show.legend = FALSE) +
  geom_vline(xintercept = 1999, color = 'red') +
  theme_light()
```



Question 2

part a: Read in data

We will use a data set in satdat.csv. Here is a description of the variables in sat:

- State – Name of state
- Region – part of US that state is in, using Census categories
- Expenditure – Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)

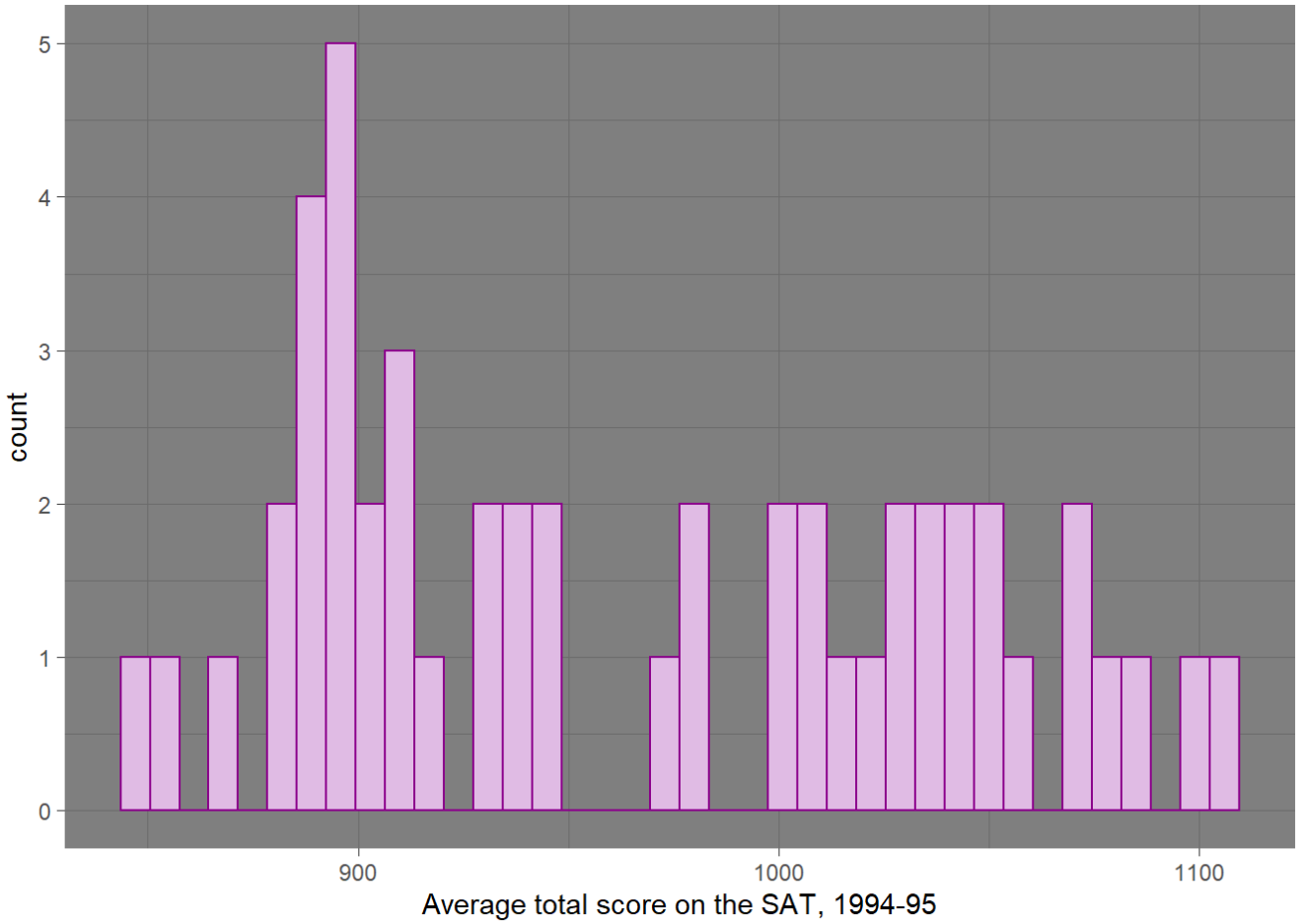
- PT.ratio – Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994
- AveSalary – Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)
- PercTaking – Percentage of all eligible students taking the SAT, 1994-95
- SATV – Average verbal SAT score, 1994-95
- SATM – Average math SAT score, 1994-95
- SATTot – Average total score on the SAT, 1994-95

```
#  Put the satdata.csv file in your course folder and read in the data set:  satdat
a.csv, using this method:

sat <- read.csv('satdata.csv', stringsAsFactors = TRUE)

#  Do head(sat)
head(sat)
```

```
##       State Region Expenditure PT.ratio AveSalary PercTaking SATV SATM SATTot
## 1  Illinois     MW       6.136     17.3    39.431         13  488  560   1048
## 2   Indiana     MW       5.826     17.5    36.785         58  415  467    882
## 3      Iowa     MW       5.483     15.8    31.511          5  516  583   1099
## 4    Kansas     MW       5.817     15.1    34.652          9  503  557   1060
## 5  Michigan     MW       6.994     20.1    41.895         11  484  549   1033
## 6 Minnesota     MW       6.000     17.5    35.948          9  506  579   1085
```

part b: SAT Total score

```
#  We will focus on the SATTot variable. Using ggplot, make a histogram of it. Make
it look nice by filling bars with a light color (e.g., yellow), and  outlining them
with a dark color (e.g., blue).  Try changing the bins.
ggplot(data = sat,
       mapping = aes( x = SATTot )) +
  geom_histogram(fill = '#E0BBE4', color = '#8b008b', binwidth = 7) +
  theme_dark() +
  labs(x = "Average total score on the SAT, 1994-95" )
```

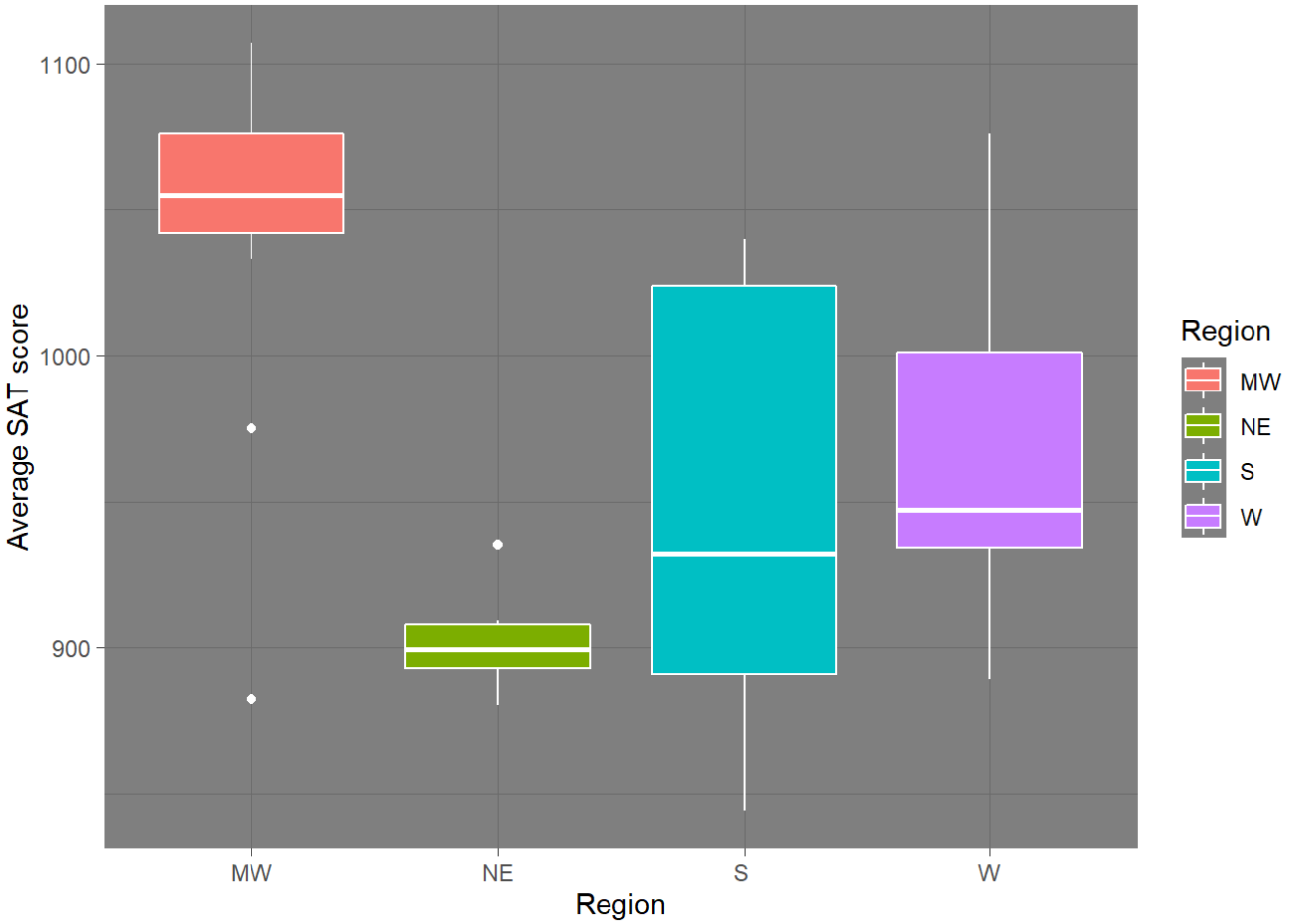


part c: SATTot by Region

< Make the graph in the chunk below, then describe it here. Which regions have higher scores? Lower scores?>

As seen below, the region with the highest scores is the Midwest. While the Northeast has lower scores as the boxplot's median indicates. In these two regions, there does not seem to be a great variety which could indicate that in general the students of those states are better and worse prepared respectively. In the South we can see that there is a very big variation of scores which is very interesting.

```
#  To see if scores vary across regions of the US, do a boxplot of SATTot by Regio
n. Use color.
ggplot(data = sat,
       mapping = aes( x = Region , y = SATTot, fill = Region )) +
  geom_boxplot(color = "white") +
  theme_dark() +
  labs(y = "Average SAT score", x = "Region" )
```

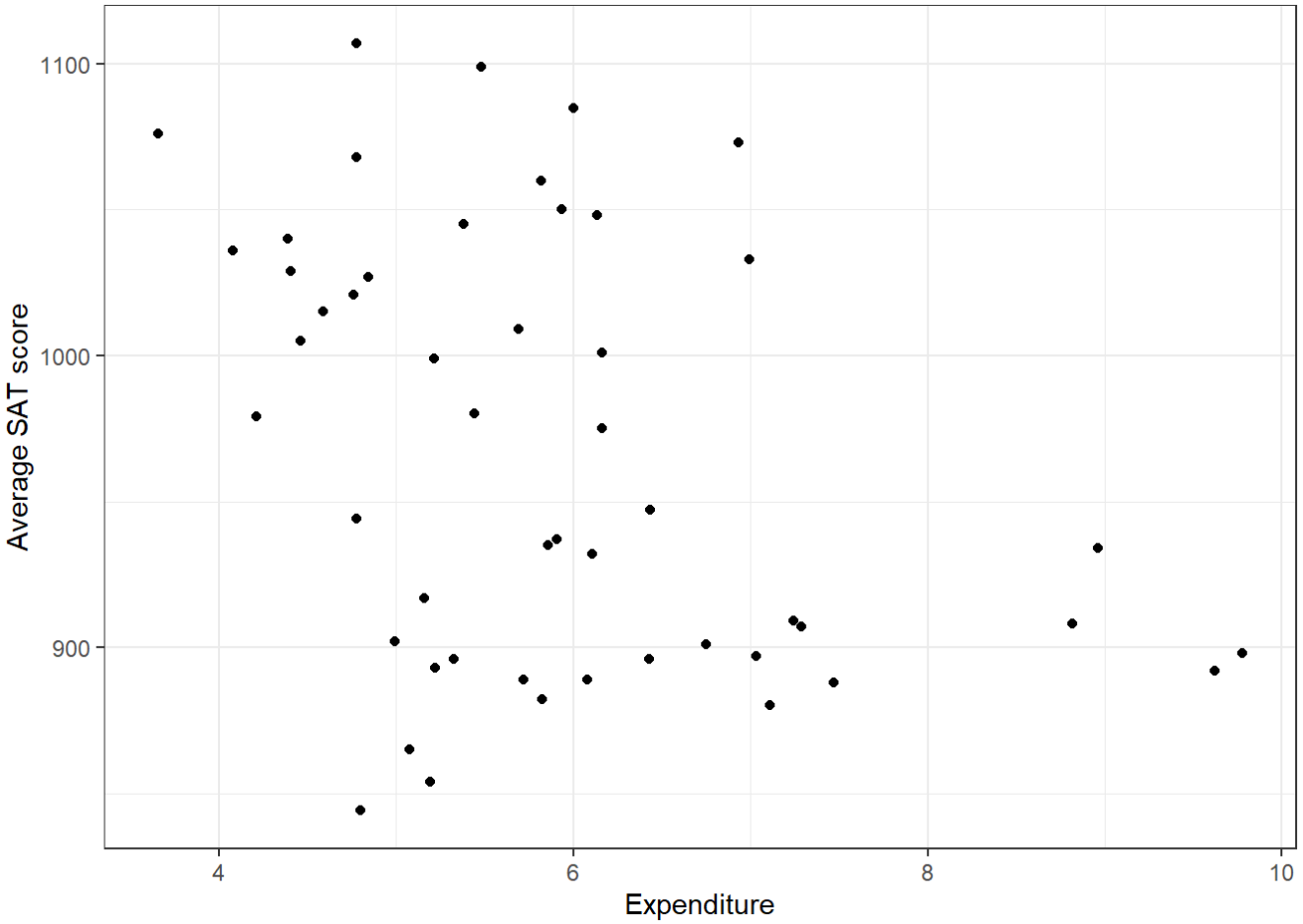


part d: Why do regions differ so much?

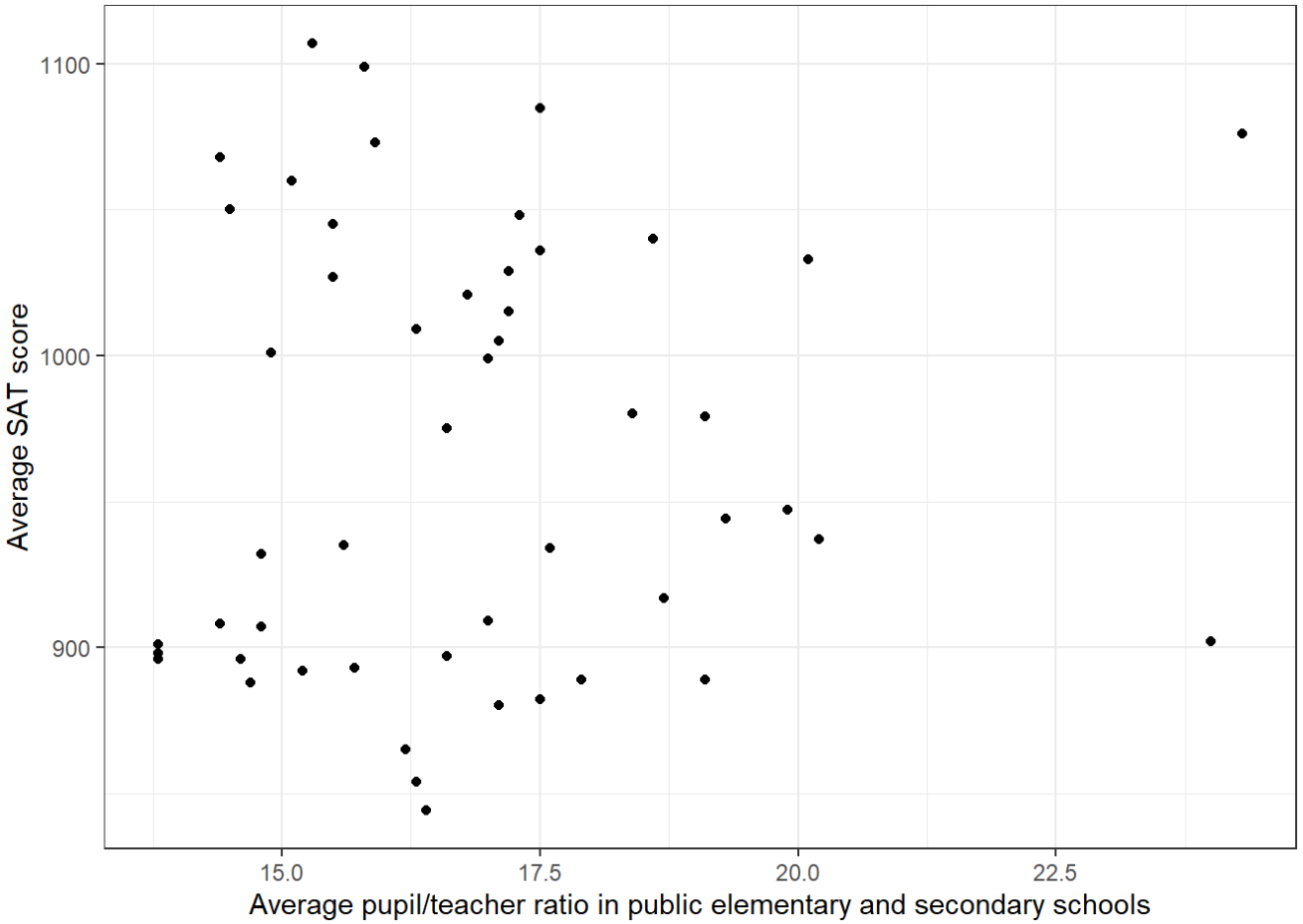
< Make the graphs below, then describe them here: Which of the variables shows the strongest relationship with SATTot? Briefly describe the plot of the strongest relationship, in terms of the variables.>

Looking at the graphs, we can clearly identify the variable who has the strongest relationship with the total average score in the SATs (SATTot). Percentage of all eligible students taking the SAT has negative, moderate and linear relationship with the first variable.

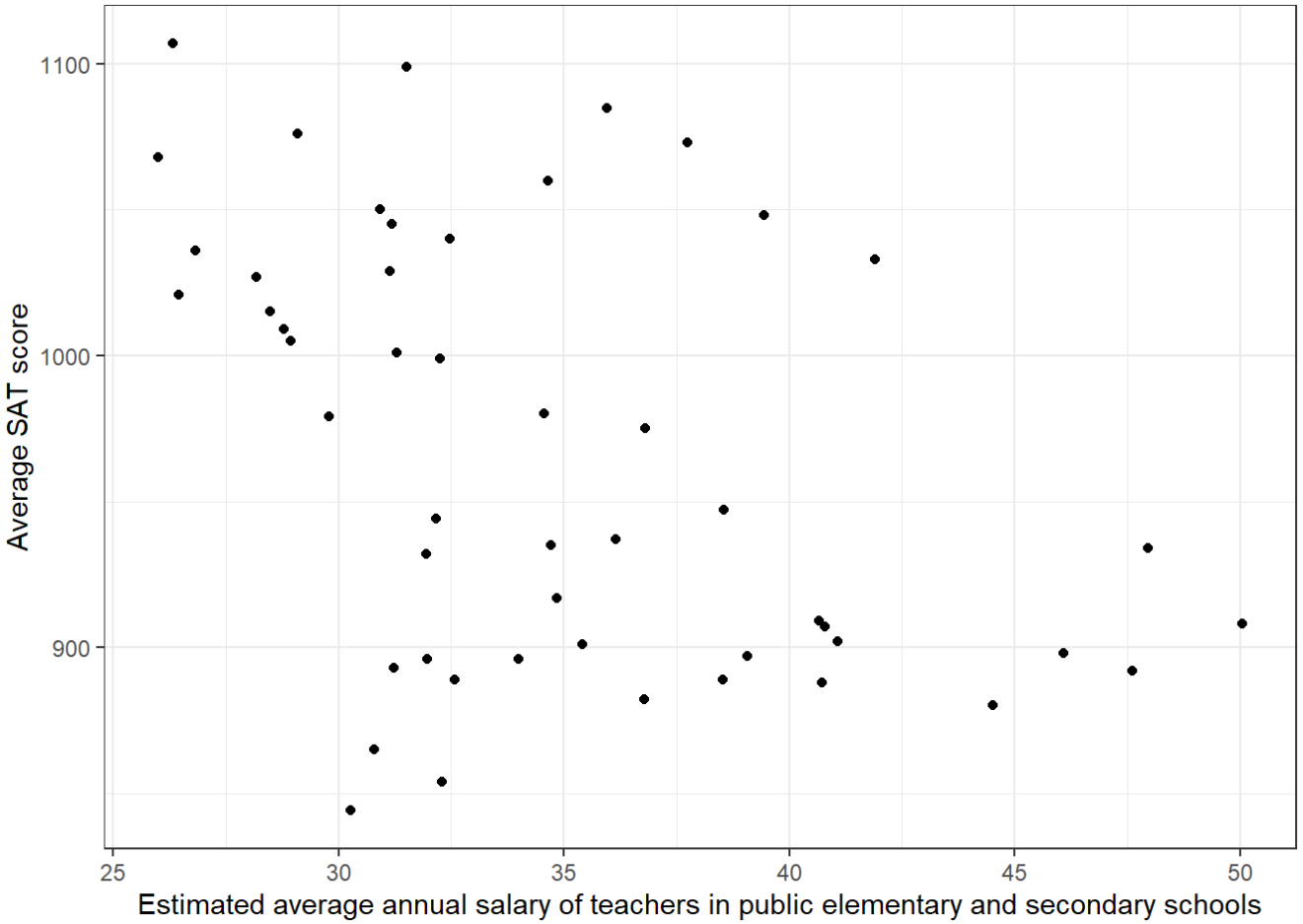
```
#  Why do the regions differ so much in scores? (And what is wrong with the Northeast? :-/ ) To find out, first try this:
#  Using ggplot, create scatterplots to see which of the variables Expenditure, PT.ratio, AveSalary, and PercTaking is most correlated with SATTot.  That is, let y = SATTot, and make four scatterplots, letting each variable be x.
ggplot(data = sat,
       mapping = aes( x = Expenditure , y = SATTot )) +
  geom_point() +
  theme_bw() +
  labs(y = "Average SAT score", x = "Expenditure" )
```



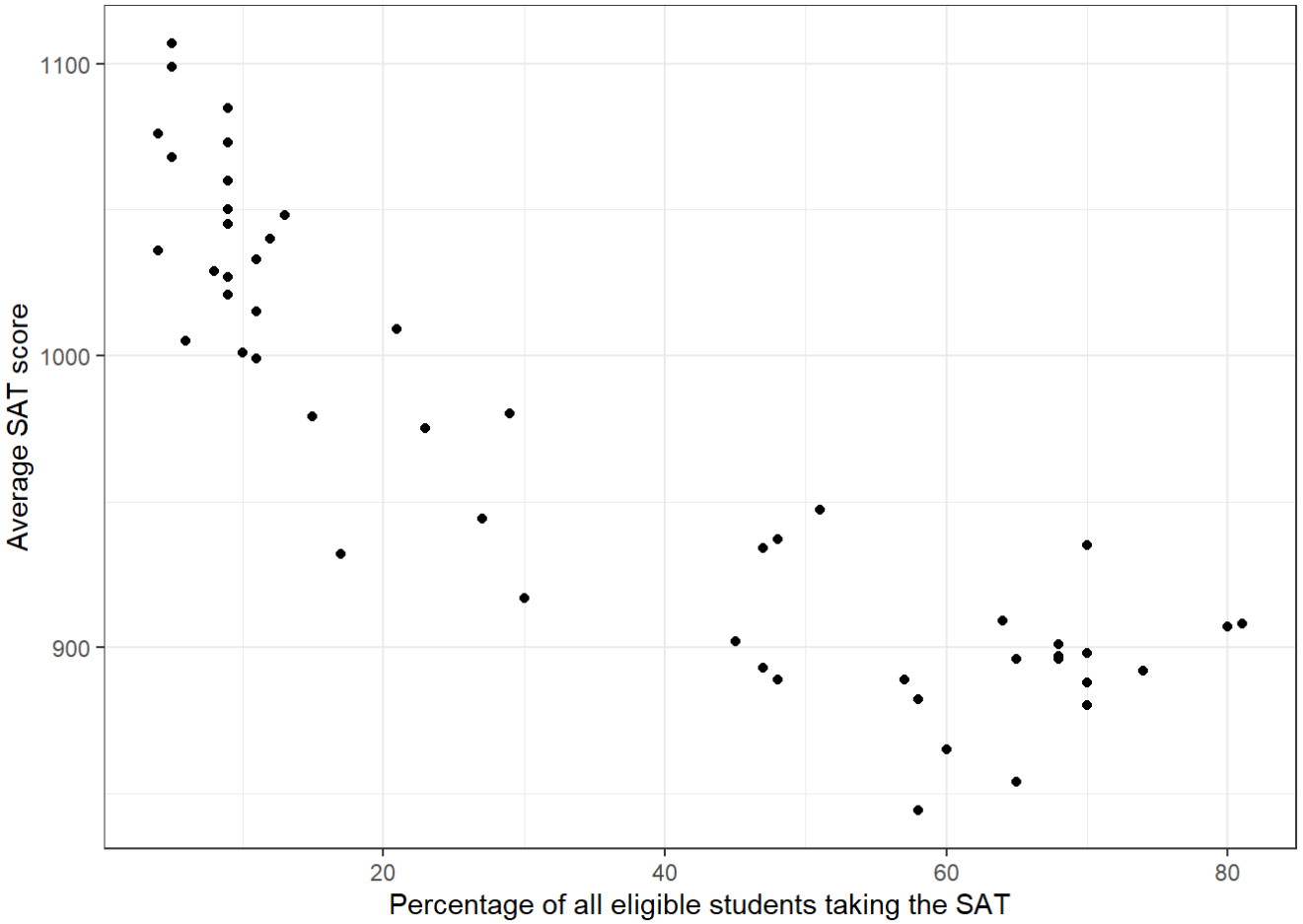

```
ggplot(data = sat,
       mapping = aes( x = PT.ratio , y = SATTot )) +
  geom_point() +
  theme_bw() +
  labs(y = "Average SAT score", x = "Average pupil/teacher ratio in public elementa
ry and secondary schools" )
```




```
ggplot(data = sat,
       mapping = aes( x = AveSalary , y = SATTot )) +
  geom_point() +
  theme_bw() +
  labs(y = "Average SAT score", x = "Estimated average annual salary of teachers in
public elementary and secondary schools" )
```



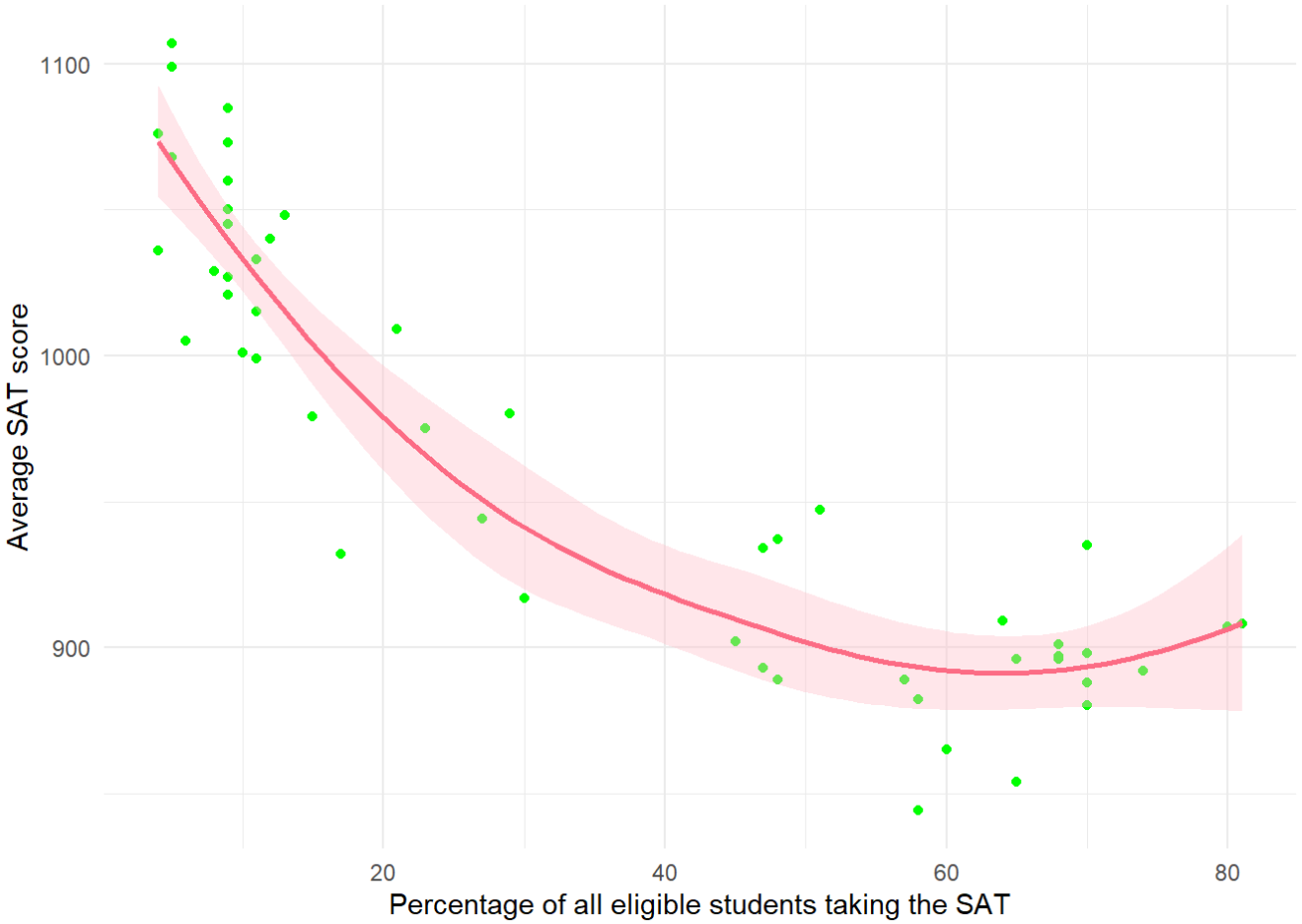

```
ggplot(data = sat,
       mapping = aes( x = PercTaking , y = SATTot )) +
  geom_point() +
    theme_bw() +
  labs(y = "Average SAT score", x = "Percentage of all eligible students taking the
SAT" )
```



part e: Strongest Relationship

< Describe the relationship that you see in the plot below. Why might this relationship exist?>

```
#  Redo the plot with that strongest relating variable, adding a smoothed line (or
curve) to the plot.
ggplot(data = sat,
       mapping = aes( x = PercTaking , y = SATTot )) +
  geom_point(color = 'green') +
  stat_smooth(color = '#fc6c85', fill = 'pink') +
    theme_minimal() +
  labs(y = "Average SAT score", x = "Percentage of all eligible students taking the
SAT" )
```

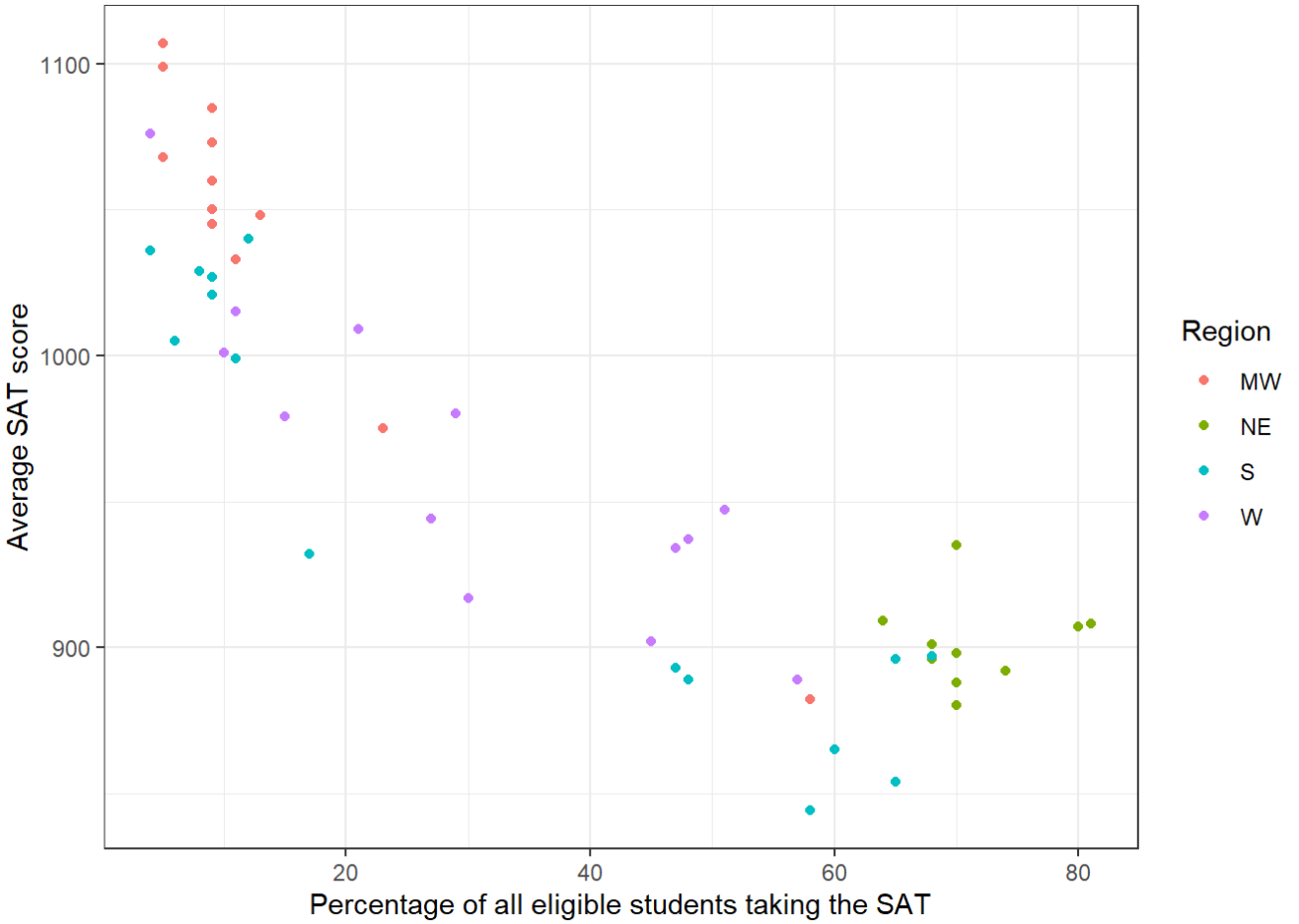


part f: Does it differ by Region?

< Describe your plot here – Is there a tendency for certain regions to fall in a certain part of the plot? >

Looking at said graph, we can see that in the Midwest, there was a much larger quantity of students eligible to take the SATs who had a high score than those who had a low score. However, the Northeast has the opposite tendency. A possible reason might be that the States in the Northeast are more lenient towards who can take part in the SATs, unlike the Midwest.

```
#  Now, redo the scatterplot of that one variable  again, this time WITHOUT the smo
othed line, but letting the points be colored by the Region of the states.
ggplot(data = sat,
       mapping = aes( x = PercTaking , y = SATTot, color = Region )) +
  geom_point() +
    theme_bw() +
  labs(y = "Average SAT score", x = "Percentage of all eligible students taking the
SAT" )
```

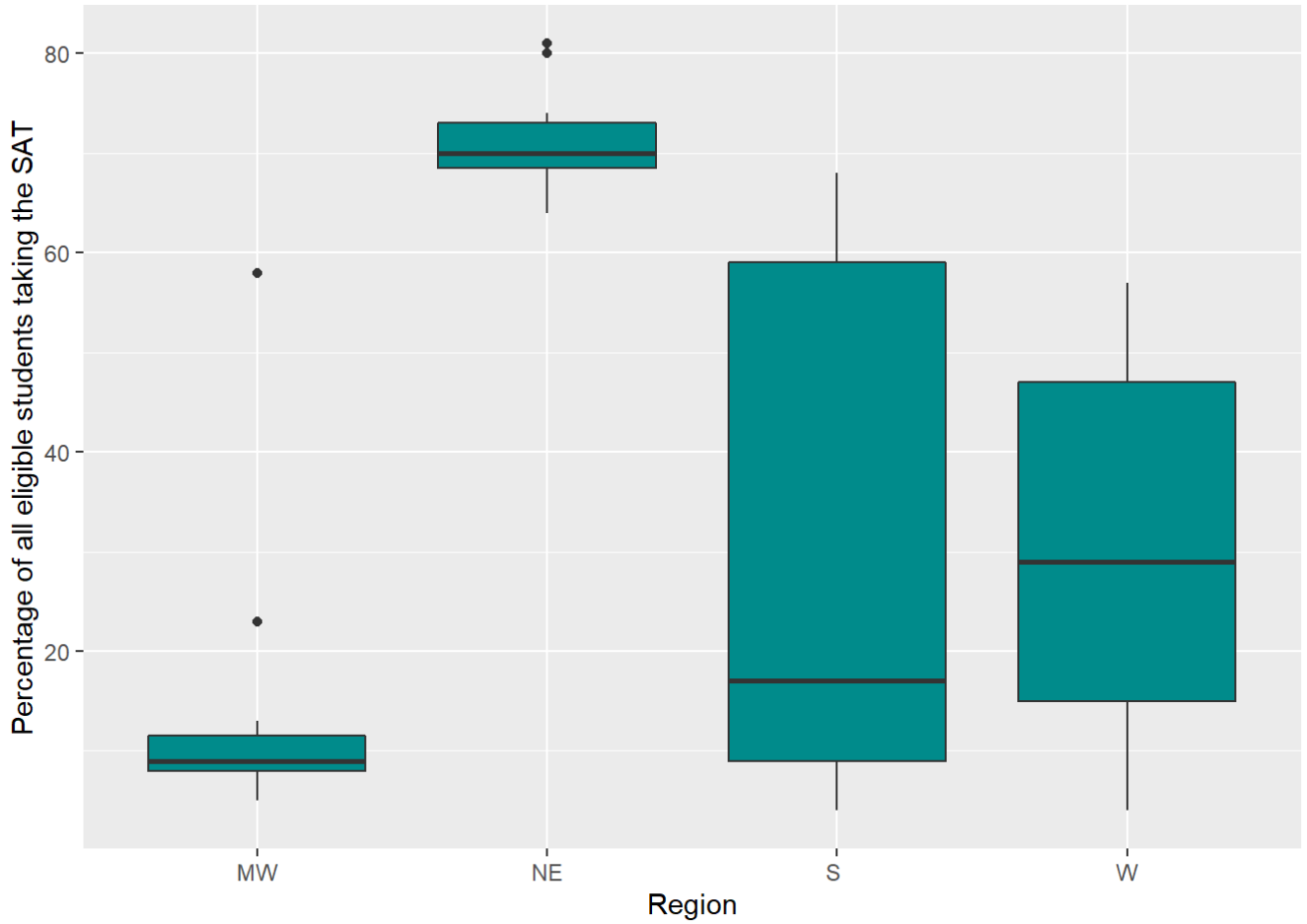


part g: Boxplots by Region

< To check the tendency, you will make a boxplot of your strongest variable by Region. Describe here what you see>

As we theorized previously, their is a dramatic difference between the quantity of students eligible to take the SATs in both regions. The Northeastern regions have 7 times more percentage of students elegible than the Midwestern.

```
# make a boxplot of your strongest variable by Region.
ggplot(data = sat,
       mapping = aes( x = Region , y = PercTaking)) +
  geom_boxplot(fill = "#008b8b") +
    theme_gray() +
  labs(x = "Region", y = "Percentage of all eligible students taking the SAT" )
```



part h:

< Summarize what you learned about why test scores varied across the US as they did. What is your opinion on whether SAT tests should be used for college admission? >

During this data analysis I have learned that the average test scores in the SATs per region does not necessarily mean that students were less prepared, but that depending on the region, more or less students are elegible to take the SATs.

As a foreign student that has only heard of the SATs through TV shows, I do not have a formed opinion on whether SAT testing is appropriate or not. However, I do think that all regions should have the same percentage of students elegible to take the SATs. I also do not quite understand why in some places more students are elegible than others, is there different requirements according to the state?

Question 3

part a: Read in Survey data

```
#    Put surveyB_F21.csv in your class folder, and read it like this
# sb <- read.csv('surveyB_F21.csv', stringsAsFactors = TRUE)
sb <- read.csv('surveyB_F21.csv', stringsAsFactors = TRUE)

#   Do a summary() of the data frame.
summary(sb)
```

```
##  ï..Response_id           Form               Year
##  Min.   :1166955   Long1000 :112   0 Non-Degree: 4
##  1st Qu.:1170999   Short3000: 88   1st year    :22
##  Median :1171793                   2 Sophomore :27
##  Mean   :1171495                   3 Junior    :55
##  3rd Qu.:1172461                   4 Senior    :85
##  Max.   :1174326                   5 Graduate  : 7
##                         Drive          PAS                 Relationship
##                            : 2   Oppose : 20                    :  2
##  1 Way Worse than Average : 4   Support:180   No                :113
##  2 Worse than Average     :16                 Sort of/ Not sure: 14
##  3 Average                :66                 Yes               : 71
##  4 Better than Average    :91
##  5 Way Better than Average:21
```

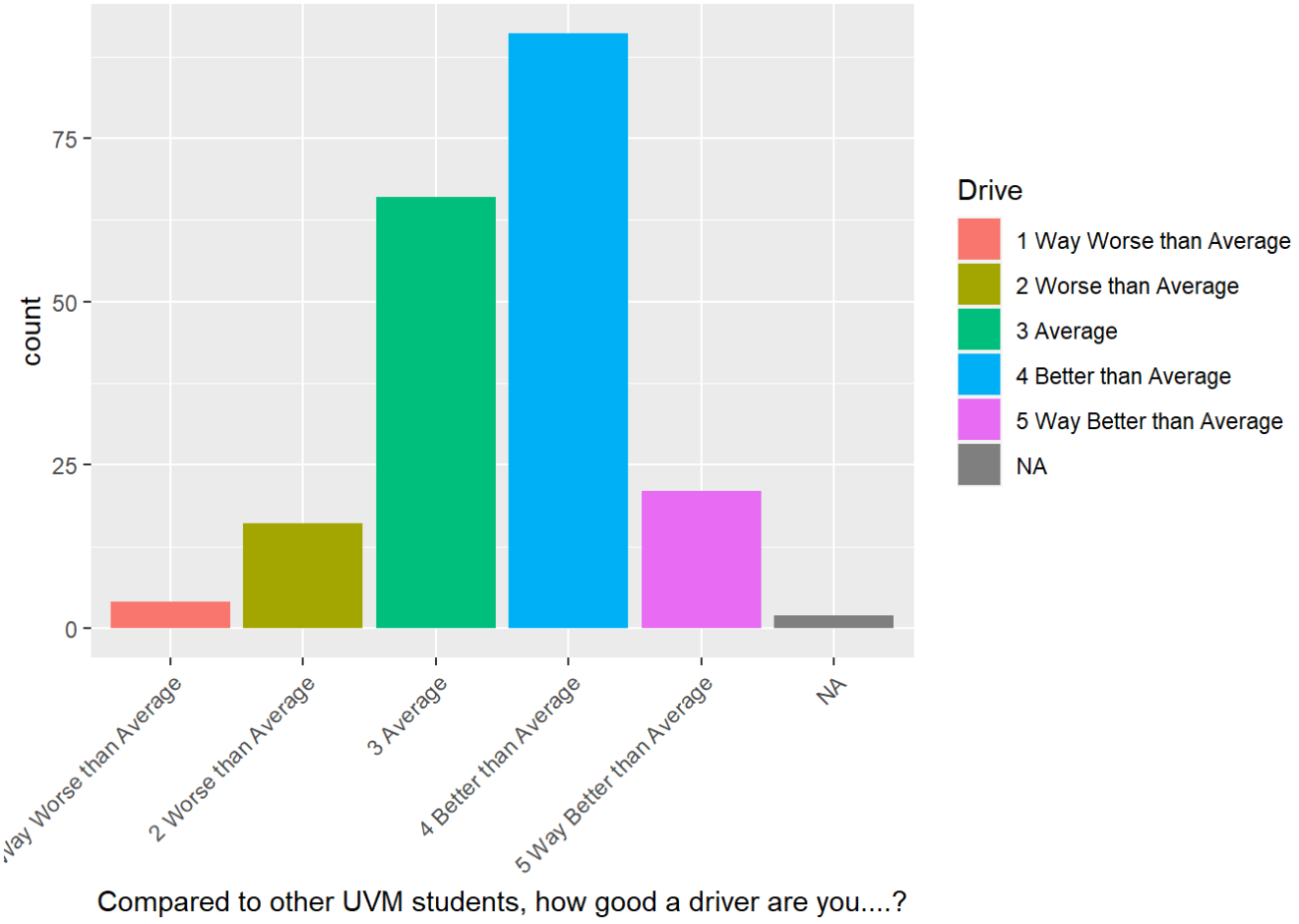
```
#  Notice the vectors Drive and Relationship have some blank values. We want these blanks to be read as missing value by R.   To fix this, redo your read.csv, adding the argument below.  Redo your summary, and notice that it will turn the blanks int o NAs:
sb <- read.csv('surveyB_F21.csv', stringsAsFactors = TRUE, na.strings = '')
summary(sb)
```

```
##  ï..Response_id           Form               Year
##  Min.   :1166955   Long1000 :112   0 Non-Degree: 4
##  1st Qu.:1170999   Short3000: 88   1st year    :22
##  Median :1171793                   2 Sophomore :27
##  Mean   :1171495                   3 Junior    :55
##  3rd Qu.:1172461                   4 Senior    :85
##  Max.   :1174326                   5 Graduate  : 7
##                         Drive          PAS                 Relationship
##  1 Way Worse than Average : 4   Oppose : 20   No                :113
##  2 Worse than Average     :16   Support:180   Sort of/ Not sure: 14
##  3 Average                :66                 Yes               : 71
##  4 Better than Average    :91                 NA's              :  2
##  5 Way Better than Average:21
##  NA's                     : 2
```

part b: Graph the Drive variable

< After you graph, describe here: Which response is the most popular? The least popular (besides NA)? Does anything surprise you about the results? Explain.> The most popular response is “Better than average” and the least is “Way worse than average”. These results do not surprise me as a lot of people hold themselves at high esteem, especially considering their driving skills.

```
#  Using ggplot, make a bar graph of just the variable Drive.  The data comes from
our class survey, which asked the question:

#  "Compared to other UVM students, how good a driver are you....? "

#  Let the bars be filled with different colors.  Note that the bar with people who
did not answer the question is still there, though it is correctly labeled NA.  We
will talk about how to remove it later...

#  The x-axis labels are hard to read, because they are long.
#  Add this layer to your code to make them look better:
# theme(axis.text.x = element_text(hjust=1, angle = 45))
ggplot(data = sb,
       mapping = aes( x = Drive, fill= Drive )) +
  geom_bar() +
    theme_gray() +
  theme(axis.text.x = element_text(hjust=1, angle = 45)) +
  labs(x = "Compared to other UVM students, how good a driver are you....?")
```



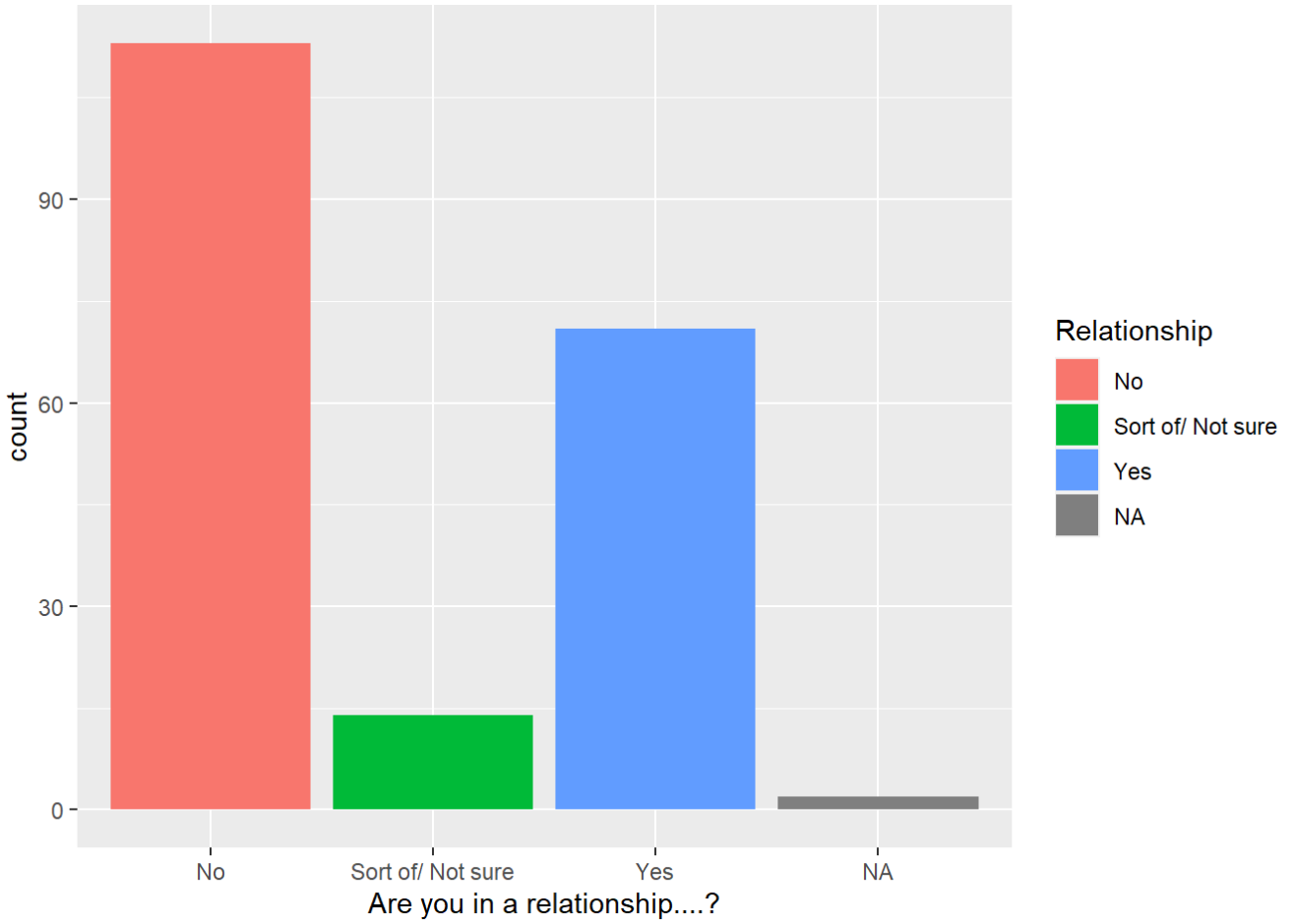
part c: Graph the Relationship variable

< After you graph, describe here: Which response is the most popular? The least popular (besides NA)? Does anything surprise you about the results? Explain.>

In this graph we can see that most UVM students have answered that they are not in a relationship. The least popular response is not the opposite but instead “Sort of, not sure”. This means that UVM students are generally decisive and do not string people along. The answers surprise me because a friend told me that

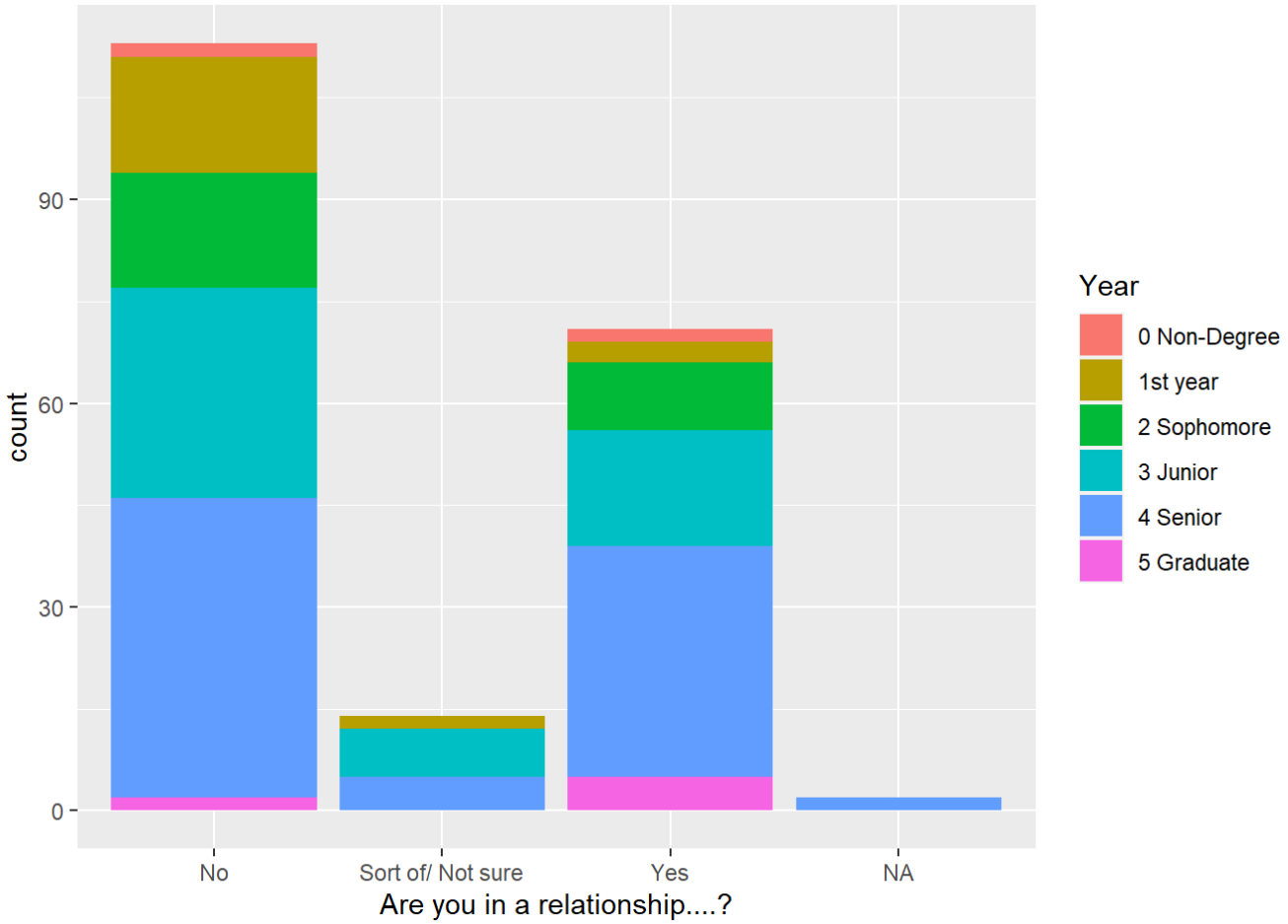
most people here were in a relationship and because in Spain a lot of my friends would reply “Sort of/not sure”.

```
#  Using ggplot, make a bar graph of just the variable Relationship.
#  Let the bars be filled with different colors.
ggplot(data = sb,
       mapping = aes( x = Relationship, fill = Relationship)) +
  geom_bar() +
    theme_gray() +
  labs(x = "Are you in a relationship....?")
```



part d: Bar graph of Relationship by Year in School.

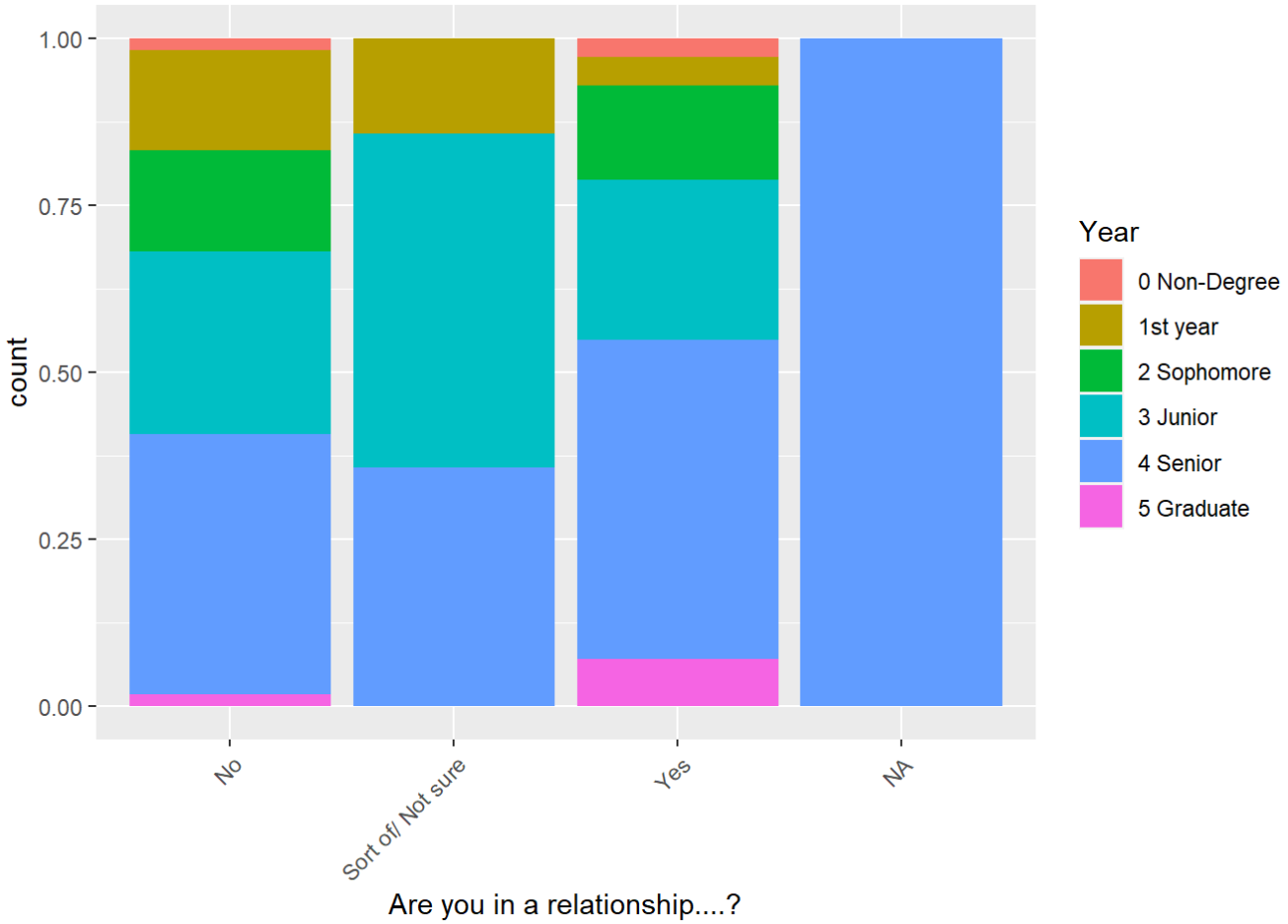
```
#  Add the fill = Year aesthetic...
ggplot(data = sb,
       mapping = aes( x = Relationship, fill = Year)) +
  geom_bar() +
    theme_gray() +
  labs(x = "Are you in a relationship....?")
```



part e: Relationship by Year with position = 'fill'

< Describe here: Does the response appear to differ by year in school? (Again, ignore the NA's for now) Explain how. Do any of the differences surprise you? > Looking at the percentages displayed we can see that the answers do vary by year. According to each column, more seniors have replied in general and the biggest repliers for "Sort of/Not sure" are Juniors. Furthermore, less first years seem to be in a relationship than upperclassmen. This data does not surprise me as it is often said that Freshmen come to college single or ending up breaking up with their long distance shortly after arriving.

```
#  To make comparisons easier, redo the graph, adding position = 'fill' in the geom
_bar()  parentheses.
ggplot(data = sb,
       mapping = aes( x = Relationship, fill = Year)) +
  geom_bar(position = 'fill') +
    theme_gray() +
  theme(axis.text.x = element_text(hjust=1, angle = 45)) +
  labs(x = "Are you in a relationship....?")
```


1.00
0.75
0.50
0.25
0.00
count
No
Sort of/ Not sure
Yes
NA
Are you in a relationship....?
Year
0 Non-Degree
1st year
2 Sophomore
3 Junior
4 Senior
5 Graduate