Education Across The United States of America

Pavel Olariu po3787

INTRODUCTION

```
library(tidyverse)
library(mosaicData)
library(carData)
States
```

```
##    region   pop SATV SATM percent dollars pay
## AL    ESC  4041  470  514       8   3.648  27
## AK    PAC   550  438  476      42   7.887  43
## AZ    MTN  3665  445  497      25   4.231  30
## AR    WSC  2351  470  511       6   3.334  23
## CA    PAC 29760  419  484      45   4.826  39
## CO    MTN  3294  456  513      28   4.809  31
## CN     NE  3287  430  471      74   7.914  43
## DE     SA   666  433  470      58   6.016  35
## DC     SA   607  409  441      68   8.210  39
## FL     SA 12938  418  466      44   5.154  30
## GA     SA  6478  401  443      57   4.860  29
## HI    PAC  1108  404  481      52   5.008  32
## ID    MTN  1007  466  502      17   3.200  25
## IL    ENC 11431  466  528      16   5.062  34
##  [ reached 'max' / getOption("max.print") -- omitted 37 rows ]
```

```
SAT
```

```
##          state expend ratio salary frac verbal math  sat
## 1      Alabama  4.405  17.2 31.144    8    491  538 1029
## 2       Alaska  8.963  17.6 47.951   47    445  489  934
## 3      Arizona  4.778  19.3 32.175   27    448  496  944
## 4     Arkansas  4.459  17.1 28.934    6    482  523 1005
## 5   California  4.992  24.0 41.078   45    417  485  902
## 6     Colorado  5.443  18.4 34.571   29    462  518  980
## 7  Connecticut  8.817  14.4 50.045   81    431  477  908
## 8     Delaware  7.030  16.6 39.076   68    429  468  897
## 9      Florida  5.718  19.1 32.588   48    420  469  889
## 10     Georgia  5.193  16.3 32.291   65    406  448  854
## 11      Hawaii  6.078  17.9 38.518   57    407  482  889
## 12       Idaho  4.210  19.1 29.783   15    468  511  979
##  [ reached 'max' / getOption("max.print") -- omitted 38 rows ]
```

```
#install packages that allow library use and opened datasets for viewing and situating
```

The datasets in this analysis both pertain to educational statistics of all 50 states present in The United States of America. One dataset uses primary and secondary school information from the 1994-1995 school year, and the other from 1992 (which can be assumed to be from the 1991-1992 school year). This allows for two sets of different variables, classified by year, for the qualities of the datasets that are common, such as the verbal SAT score (verbal[year]), math SAT score (math[year]), total SAT score (sat[year]), salary of teachers in thousands of dollars (salary[year]), expenditure per individual pupil in thousands of dollars(pupilexpense[year]), and the percentage of graduating students who had taken the SAT (percent[year]). The variables that will be considered constant across both years will be the ratio of pupils to teachers (ratio), the U.S. census region of the state (region), and the population of the states in thousands of people (pop). The region abbreviation meanings are: ENC, East North Central; ESC, East South Central; MA, Mid-Atlantic; MTN, Mountain; NE, New England; PAC, Pacific; SA, South Atlantic;WNC, West North Central; WSC, West South Central.These are considered constant as either one or the other dataset does not mirror the variable, and so cannot be compared across year, but can be used as an aid in the general relation of the other variables. These datasets were taken from two packages already present in R, called “States” and “SAT”, which were created with data taken from the “Bureau of the Census” and the “Journal of Statistics Education”. These datasets are interesting to me because I like seeing how states vary in certain features, and which are the most educated and value their education and students the most. I think that there will be a correlation between population/region and the scores/money spent of education. I think that in the northern states there will be higher scores and more money spent on education because of the democratic leanings and culture of those states. I also think that states with lower populations will score higher and spend more money on education because they have less students to worry about and can focus on them more.

TIDYING

```
states1992 <- States[-c(9),]
#deletes row describing Washington D.C. as this is not really a state and not included in the other dataset.
states1992renamed <- states1992%>%
  rename(
    verbal1992 = SATV,
    math1992 = SATM,
    percent1992 = percent,
    pupilexpense1992 = dollars,
    salary1992 = pay
  )
#renamed variables for first set to represent the associated year in the variable name for easier comparison
states1995renamed <- SAT%>%
  rename(
    pupilexpense1995 = expend,
    salary1995 = salary,
    percent1995 = frac,
    verbal1995 = verbal,
    math1995 = math,
    sat1995 = sat
  )
#renamed variables for other dataset
states1992renamedv2 <- states1992renamed%>%
  mutate(sat1992 = math1992 + verbal1992)
#created a new variable for total sat score by adding verbal and math scores together
states1992renamedv3 <- states1992renamedv2%>%
  mutate(state = states1995renamed$state)
#added a variable of states in 1992 set because the state were being used as the row names instead of its variable, had to d
o it for easier joining.
states1992renamedv3
```

```
##    region   pop verbal1992 math1992 percent1992 pupilexpense1992 salary1992
## AL    ESC  4041        470      514           8             3.648         27
## AK    PAC   550        438      476          42             7.887         43
## AZ    MTN  3665        445      497          25             4.231         30
## AR    WSC  2351        470      511           6             3.334         23
## CA    PAC 29760        419      484          45             4.826         39
## CO    MTN  3294        456      513          28             4.809         31
## CN     NE  3287        430      471          74             7.914         43
## DE     SA   666        433      470          58             6.016         35
## FL     SA 12938        418      466          44             5.154         30
## GA     SA  6478        401      443          57             4.860         29
## HI    PAC  1108        404      481          52             5.008         32
##    sat1992       state
## AL     984     Alabama
## AK     914      Alaska
## AZ     942     Arizona
## AR     981    Arkansas
## CA     903  California
## CO     969    Colorado
## CN     901 Connecticut
## DE     903    Delaware
## FL     884     Florida
## GA     844     Georgia
## HI     885      Hawaii
##  [ reached 'max' / getOption("max.print") -- omitted 39 rows ]
```

```
states1995renamed
```

```
##          state pupilexpense1995 ratio salary1995 percent1995 verbal1995
## 1      Alabama            4.405  17.2     31.144           8        491
## 2       Alaska            8.963  17.6     47.951          47        445
## 3      Arizona            4.778  19.3     32.175          27        448
## 4     Arkansas            4.459  17.1     28.934           6        482
## 5   California            4.992  24.0     41.078          45        417
## 6     Colorado            5.443  18.4     34.571          29        462
## 7  Connecticut            8.817  14.4     50.045          81        431
## 8     Delaware            7.030  16.6     39.076          68        429
## 9      Florida            5.718  19.1     32.588          48        420
## 10     Georgia            5.193  16.3     32.291          65        406
## 11      Hawaii            6.078  17.9     38.518          57        407
## 12       Idaho            4.210  19.1     29.783          15        468
##    math1995 sat1995
## 1       538    1029
## 2       489     934
## 3       496     944
## 4       523    1005
## 5       485     902
## 6       518     980
## 7       477     908
## 8       468     897
## 9       469     889
## 10      448     854
## 11      482     889
## 12      511     979
##  [ reached 'max' / getOption("max.print") -- omitted 38 rows ]
```

The datesets were similar in a few variables, so these variables were changed to reflect the year they represent. Row containing Washington D.C. was deleted from one dataset as it was not included in the other. Common variables between the two were renamed to reflect the year they are from, as the educational statistics are from two different years. A total SAT score variable was created in one of the datasets to match the other. This was done by adding the verbal and math scores together. An explicit variable listing the names of the states was added to one of the datasets as they were the row names previously, and this would have prohibited simple joining of the datasets.

JOINING

```
combinedstates <- states1995renamed%>%
  full_join(states1992renamedv3, by = "state")
combinedstates
```

```
##        state pupilexpense1995 ratio salary1995 percent1995 verbal1995 math1995
## 1    Alabama            4.405  17.2     31.144           8        491      538
## 2     Alaska            8.963  17.6     47.951          47        445      489
## 3    Arizona            4.778  19.3     32.175          27        448      496
## 4   Arkansas            4.459  17.1     28.934           6        482      523
## 5 California            4.992  24.0     41.078          45        417      485
## 6   Colorado            5.443  18.4     34.571          29        462      518
##   sat1995 region   pop verbal1992 math1992 percent1992 pupilexpense1992
## 1    1029    ESC  4041        470      514           8            3.648
## 2     934    PAC   550        438      476          42            7.887
## 3     944    MTN  3665        445      497          25            4.231
## 4    1005    WSC  2351        470      511           6            3.334
## 5     902    PAC 29760        419      484          45            4.826
## 6     980    MTN  3294        456      513          28            4.809
##   salary1992 sat1992
## 1         27     984
## 2         43     914
## 3         30     942
## 4         23     981
## 5         39     903
## 6         31     969
##  [ reached 'max' / getOption("max.print") -- omitted 44 rows ]
```

```
#joined both sets using full join to ensure all columns and rows were kept, joined with the key variable designated as 'stat
es'
```

The joined dataset incorporates all the variables present in the two individual datasets, and uses the states as a base for joining. No problems were encountered in joining the two as the states were in identical alphabetical order, and there were exactly 50 observations in each dataset, corresponding to the 50 states.

SUMMARY STATISTICS

```
library(kableExtra)
#a useful package that can help with tables
combinedstates <- combinedstates%>%
  mutate(averagesat = (sat1992 + sat1995)/2)%>%
  mutate(averagepupilexpense = (pupilexpense1992 + pupilexpense1995)/2)%>%
  mutate(averagesalary = (salary1992 + salary1995)/2)%>%
  mutate(averagepercent = (percent1992 + percent1995)/2)%>%
  mutate(averagesatverbal = (verbal1992 + verbal1995)/2)%>%
  mutate(averagesatmath = (math1992 + math1995)/2)
#created averages for all the variables that were common across both years
#combinedstates_withpercents <- combinedstates%>%
#  mutate(pctchange_sat = (sat1992 + sat1995)*100)%>%
#  mutate(pctchange_pupilexpense = (pupilexpense1995/pupilexpense1992 - 1)*100)%>%
#  mutate(pctchange_salary = (salary1995/salary1992 - 1)*100)%>%
#  mutate(pctchange_percent = (percent1995/percent1992 - 1)*100)%>%
#  mutate(pctchange_satverbal = (verbal1995/verbal1992 - 1)*100)%>%
#  mutate(pctchange_satmath = (math1995/math1992 - 1)*100)%>%
#  mutate(pctchange_averagesat = (sat1995/sat1992 - 1)*100)
#created a percent change variable that shows the change from 1992 to 1995 of the variables(talen out caused too many proble
ms)
combinedstates %>%
  group_by(region)%>%
  summarise(mean_ratio = mean(ratio),mean_averagesat = mean(averagesat),mean_averagepupilexpense = mean(averagepupilexpens
e),mean_averagesalary = mean(averagesalary),mean_averagepercent = mean(averagepercent), mean_averagesatverbal = mean(average
satverbal),mean_averagesatmath = mean(averagesatmath), mean_averagesat = mean(averagesat), mean_pop = mean(pop))%>%
  kbl%>%
  kable_styling()
```

region	mean_ratio	mean_averagesat	mean_averagepupilexpense	mean_averagesalary	mean_averagepercent	mean_averagesatverbal	mea
ENC	17.48000	980.8000	5.900300	36.16590	22.900000	460.2000	
ESC	17.57500	1010.7500	4.144625	28.83700	8.625000	483.0000	
MA	15.36667	887.6667	8.449833	42.36817	69.500000	418.0000	
MTN	18.52500	982.6875	4.647938	29.56312	18.187500	465.8750	
NE	14.51667	902.5000	6.750167	36.63892	69.500000	430.0833	
PAC	19.92000	915.7000	6.043200	38.12530	48.000000	429.7000	
SA	16.37500	883.0000	5.515187	32.41369	52.812500	420.0625	
WNC	15.44286	1058.2857	4.963357	28.96736	8.071429	498.7143	
WSC	16.27500	974.3750	4.326625	26.97375	17.125000	464.3750	

```
#  calculated the mean for all variables and grouped them by the region the states were in and then put these in a pretty ta
ble using kable
combinedstates %>%
  group_by(region)%>%
  summarise(sd_ratio = sd(ratio),sd_averagesat = sd(averagesat),sd_averagepupilexpense = sd(averagepupilexpense),sd_averages
alary = sd(averagesalary),sd_averagepercent = sd(averagepercent), sd_averagesatverbal = sd(averagesatverbal),sd_averagesatma
th = sd(averagesatmath), sd_averagesat = sd(averagesat), sd_pop = sd(pop))%>%
  kbl%>%
  kable_styling()
```

region	sd_ratio	sd_averagesat	sd_averagepupilexpense	sd_averagesalary	sd_averagepercent	sd_averagesatverbal	sd_averagesatmath
ENC	1.5943651	66.910948	0.4012098	2.317923	19.122631	29.829935	37.184338
ESC	0.7135592	11.891874	0.4670140	2.379081	3.497022	6.620675	5.299371
MA	1.6563011	6.525591	1.4246432	2.292805	2.500000	2.179450	6.934215
MTN	2.7819572	41.184201	0.8193933	2.647333	8.717542	20.928023	20.791374
NE	0.6823977	15.996875	0.9578810	5.934275	5.830952	6.909535	9.265078
PAC	2.5587106	20.271902	1.3751506	4.653343	4.227884	17.016903	4.500000
SA	1.3905292	32.841611	0.7533109	3.833027	15.836074	16.140981	16.901897
WNC	1.0390014	26.044605	0.5432636	3.719172	2.805182	11.228153	15.476941
WSC	0.7932003	61.209170	0.3427347	1.761797	18.304713	32.438082	28.818397

```
#created a table with a summary of standard deviations for all variables grouped by region
combinedstates %>%
  group_by(region)%>%
  summarise(quantile_ratio = quantile(ratio, probs = c(0.5)),quantile_averagesat = quantile(averagesat, probs = c(0.5)),quan
tile_averagepupilexpense = quantile(averagepupilexpense, probs = c(0.5)),quantile_averagesalary = quantile(averagesalary, pr
obs = c(0.5)),quantile_averagepercent = quantile(averagepercent, probs = c(0.5)), quantile_averagesatverbal = quantile(avera
gesatverbal, probs = c(0.5)),quantile_averagesatmath = quantile(averagesatmath, probs = c(0.5)), quantile_averagesat = quant
ile(averagesat, probs = c(0.5)), quantile_pop = quantile(pop, probs = c(0.5)))%>%
  kbl%>%
  kable_styling()
```

region	quantile_ratio	quantile_averagesat	quantile_averagepupilexpense	quantile_averagesalary	quantile_averagepercent	quantile_average
ENC	17.30	1000.50	5.90050	35.37300	14.50	
ESC	17.35	1011.25	4.03700	29.65525	9.25	
MA	15.20	887.00	9.06150	42.04350	69.50	
MTN	18.55	981.75	4.68900	28.76750	18.25	
NE	14.55	901.25	6.53200	35.80025	67.25	
PAC	19.90	924.00	5.54300	35.27750	46.00	
SA	16.35	891.00	5.38975	30.96975	59.25	
WNC	15.30	1052.00	5.15800	29.59450	9.50	
WSC	16.25	1000.00	4.34000	26.15825	9.00	

```
#created a table with quantile summary which gives value below which 50 percent of the data falls, which is basically the me
dian. grouped by region
combinedstates %>%
  group_by(region)%>%
  summarise(min_ratio = min(ratio),min_averagesat = min(averagesat),min_averagepupilexpense = min(averagepupilexpense),min_a
veragesalary = min(averagesalary),min_averagepercent = min(averagepercent), min_averagesatverbal = min(averagesatverbal),min
_averagesatmath = min(averagesatmath), min_averagesat = min(averagesat), min_pop = min(pop))%>%
  kbl%>%
  kable_styling()
```

region	min_ratio	min_averagesat	min_averagepupilexpense	min_averagesalary	min_averagepercent	min_averagesatverbal	min_averages
ENC	15.9	874.5	5.4385	34.3925	10.0	411.5	
ESC	17.0	996.5	3.7010	25.4090	4.0	475.0	
MA	13.8	881.5	6.8215	40.2550	67.0	415.5	
MTN	14.9	919.0	3.3245	27.0410	4.5	434.0	
NE	13.8	885.5	5.6815	29.9860	64.0	423.5	
PAC	17.6	887.0	4.9090	34.5755	44.5	405.5	
SA	14.6	839.0	4.5620	28.9720	16.0	399.0	
WNC	14.4	1020.0	4.2300	23.9970	5.0	484.0	
WSC	15.5	883.5	3.8965	25.9670	6.0	416.0	

```
#created a table with a summary of the minimum values for each variable, grouped by region
combinedstates %>%
  group_by(region)%>%
  summarise(max_ratio = max(ratio),max_averagesat = max(averagesat),max_averagepupilexpense = max(averagepupilexpense),max_a
veragesalary = max(averagesalary),max_averagepercent = max(averagepercent), max_averagesatverbal = max(averagesatverbal),max
_averagesatmath = max(averagesatmath), max_averagesat = max(averagesat), max_pop = max(pop))%>%
  kbl%>%
  kable_styling()
```

region	max_ratio	max_averagesat	max_averagepupilexpense	max_averagesalary	max_averagepercent	max_averagesatverbal	max_avera
ENC	20.1	1046.0	6.4380	39.9475	56.0	488.5	
ESC	18.6	1024.0	4.8035	30.6285	12.0	490.0	

region	max_ratio	max_averagesat	max_averagepupilexpense	max_averagesalary	max_averagepercent	max_averagesatverbal	max_avera
MA	17.1	894.5	9.4665	44.8060	72.0	419.5	
MTN	24.3	1053.5	5.7075	33.4180	28.5	502.5	
NE	15.6	931.5	8.3655	46.5225	77.5	443.0	
PAC	24.0	935.0	8.4250	45.4755	54.5	443.5	
SA	19.1	932.5	6.7145	39.3305	63.0	445.5	
WNC	17.5	1093.5	5.6300	34.4740	11.5	513.5	
WSC	17.1	1014.0	4.7300	29.6115	44.5	484.5	

```
#created a table with a summary of the maximum values for each variable, grouped by states.
combinedstates %>%
  group_by(region)%>%
  summarise(iqr_ratio = IQR(ratio),iqr_averagesat = IQR(averagesat),iqr_averagepupilexpense = IQR(averagepupilexpense),iqr_a
veragesalary = IQR(averagesalary),iqr_averagepercent = IQR(averagepercent), iqr_averagesatverbal = IQR(averagesatverbal),iqr
_averagesatmath = IQR(averagesatmath), iqr_averagesat = IQR(averagesat), iqr_pop = IQR(pop))%>%
  kbl%>%
  kable_styling()
```

region	iqr_ratio	iqr_averagesat	iqr_averagepupilexpense	iqr_averagesalary	iqr_averagepercent	iqr_averagesatverbal	iqr_averagesatmath
ENC	0.900	59.000	0.526500	2.314500	11.000	22.000	37.00
ESC	0.625	14.000	0.291375	2.179750	3.875	8.250	5.75
MA	1.650	6.500	1.322500	2.275500	2.500	2.000	6.75
MTN	2.175	35.375	0.899375	4.156750	14.750	16.125	24.75
NE	0.825	11.250	0.944500	5.802000	8.875	4.500	8.00
PAC	2.300	27.500	0.388000	4.780000	5.000	23.500	7.50
SA	0.850	50.125	0.808375	4.297375	8.000	24.875	24.75
WNC	0.850	31.250	0.711250	4.228500	4.750	17.500	17.50
WSC	1.225	43.125	0.278125	1.019500	9.625	20.875	22.25

```
#created a table with a summary of the interquartile range for each variable, grouped by region
combinedstates %>%
  summarise(mean_ratio = mean(ratio),mean_averagesat = mean(averagesat),mean_averagepupilexpense = mean(averagepupilexpens
e),mean_averagesalary = mean(averagesalary),mean_averagepercent = mean(averagepercent), mean_averagesatverbal = mean(average
satverbal),mean_averagesatmath = mean(averagesatmath), mean_averagesat = mean(averagesat), mean_pop = mean(pop))%>%
  kbl%>%
  kable_styling()
```

mean_ratio	mean_averagesat	mean_averagepupilexpense	mean_averagesalary	mean_averagepercent	mean_averagesatverbal	mean_avera
16.858	956.69	5.51003	32.80446	34.15	453.04	

```
combinedstates %>%
  summarise(sd_ratio = sd(ratio),sd_averagesat = sd(averagesat),sd_averagepupilexpense = sd(averagepupilexpense),sd_averages
alary = sd(averagesalary),sd_averagepercent = sd(averagepercent), sd_averagesatverbal = sd(averagesatverbal),sd_averagesatma
th = sd(averagesatmath), sd_averagesat = sd(averagesat), sd_pop = sd(pop))%>%
  kbl%>%
  kable_styling()
```

sd_ratio	sd_averagesat	sd_averagepupilexpense	sd_averagesalary	sd_averagepercent	sd_averagesatverbal	sd_averagesatmath	sd_po
2.266355	69.00672	1.330486	5.539907	25.26699	32.71738	36.84181	5459.78

```
combinedstates %>%
  summarise(quantile_ratio = quantile(ratio, probs = c(0.5)),quantile_averagesat = quantile(averagesat, probs = c(0.5)),quan
tile_averagepupilexpense = quantile(averagepupilexpense, probs = c(0.5)),quantile_averagesalary = quantile(averagesalary, pr
obs = c(0.5)),quantile_averagepercent = quantile(averagepercent, probs = c(0.5)), quantile_averagesatverbal = quantile(avera
gesatverbal, probs = c(0.5)),quantile_averagesatmath = quantile(averagesatmath, probs = c(0.5)), quantile_averagesat = quant
ile(averagesat, probs = c(0.5)), quantile_pop = quantile(pop, probs = c(0.5)))%>%
  kbl%>%
  kable_styling()
```

quantile_ratio	quantile_averagesat	quantile_averagepupilexpense	quantile_averagesalary	quantile_averagepercent	quantile_averagesatverba
16.6	939	5.4245	31.56	26.5	44

```
combinedstates %>%
  summarise(min_ratio = min(ratio),min_averagesat = min(averagesat),min_averagepupilexpense = min(averagepupilexpense),min_a
veragesalary = min(averagesalary),min_averagepercent = min(averagepercent), min_averagesatverbal = min(averagesatverbal),min
_averagesatmath = min(averagesatmath), min_averagesat = min(averagesat), min_pop = min(pop))%>%
  kbl%>%
  kable_styling()
```

min_ratio	min_averagesat	min_averagepupilexpense	min_averagesalary	min_averagepercent	min_averagesatverbal	min_averagesatmath
13.8	839	3.3245	23.997	4	399	440

```
combinedstates %>%
  summarise(max_ratio = max(ratio),max_averagesat = max(averagesat),max_averagepupilexpense = max(averagepupilexpense),max_a
veragesalary = max(averagesalary),max_averagepercent = max(averagepercent), max_averagesatverbal = max(averagesatverbal),max
_averagesatmath = max(averagesatmath), max_averagesat = max(averagesat), max_pop = max(pop))%>%
  kbl%>%
  kable_styling()
```

max_ratio	max_averagesat	max_averagepupilexpense	max_averagesalary	max_averagepercent	max_averagesatverbal	max_averagesatma
24.3	1093.5	9.4665	46.5225	77.5	513.5	58

```
combinedstates %>%
  summarise(iqr_ratio = IQR(ratio),iqr_averagesat = IQR(averagesat),iqr_averagepupilexpense = IQR(averagepupilexpense),iqr_a
veragesalary = IQR(averagesalary),iqr_averagepercent = IQR(averagepercent), iqr_averagesatverbal = IQR(averagesatverbal),iqr
_averagesatmath = IQR(averagesatmath), iqr_averagesat = IQR(averagesat), iqr_pop = IQR(pop))%>%
  kbl%>%
  kable_styling()
```

iqr_ratio	iqr_averagesat	iqr_averagepupilexpense	iqr_averagesalary	iqr_averagepercent	iqr_averagesatverbal	iqr_averagesatmath	iqr_po
2.35	116.875	1.46525	6.26025	49.625	56.75	57.375	4598.2

```
#performed same statistical measures on all variables but did not group by region, so now the statistics are calculated for
 all states.
combinedstates%>%
  group_by(state)%>%
    filter(pop>10000)%>%
  arrange(desc(averagesat))
```

```
## # A tibble: 7 x 22
## # Groups:   state [7]
##   state        pupilexpense1995 ratio salary1995 percent1995 verbal1995 math1995
##   <fct>                   <dbl> <dbl>      <dbl>       <int>      <int>    <int>
## 1 Illinois                 6.14  17.3       39.4          13        488      560
## 2 Ohio                     6.16  16.6       36.8          23        460      515
## 3 California               4.99  24         41.1          45        417      485
## 4 New York                 9.62  15.2       47.6          74        419      473
## 5 Florida                  5.72  19.1       32.6          48        420      469
## 6 Texas                    5.22  15.7       31.2          47        419      474
## 7 Pennsylvania             7.11  17.1       44.5          70        419      461
## # ... with 15 more variables: sat1995 <int>, region <fct>, pop <int>,
## #   verbal1992 <int>, math1992 <int>, percent1992 <int>,
## #   pupilexpense1992 <dbl>, salary1992 <int>, sat1992 <int>, averagesat <dbl>,
## #   averagepupilexpense <dbl>, averagesalary <dbl>, averagepercent <dbl>,
## #   averagesatverbal <dbl>, averagesatmath <dbl>
```

```
combinedstates %>%
  filter(averagesat>1000)%>%
  group_by(state)%>%
  arrange(desc(averagepupilexpense))
```

```
## # A tibble: 17 x 22
## # Groups:   state [17]
##    state       pupilexpense1995 ratio salary1995 percent1995 verbal1995 math1995
##    <fct>                  <dbl> <dbl>      <dbl>       <int>      <int>    <int>
##  1 Wisconsin               6.93  15.9       37.7           9        501      572
##  2 Michigan                6.99  20.1       41.9          11        484      549
##  3 Minnesota               6     17.5       35.9           9        506      579
##  4 Illinois                6.14  17.3       39.4          13        488      560
##  5 Kansas                  5.82  15.1       34.7           9        503      557
##  6 Iowa                    5.48  15.8       31.5           5        516      583
##  7 Nebraska                5.94  14.5       30.9           9        494      556
##  8 Missouri                5.38  15.5       31.2           9        495      550
##  9 New Mexico              4.59  17.2       28.5          11        485      530
## 10 Louisiana               4.76  16.8       26.5           9        486      535
## 11 Oklahoma                4.84  15.5       28.2           9        491      536
## 12 South Dako~             4.78  14.4       26.0           5        505      563
## 13 North Dako~             4.78  15.3       26.3           5        515      592
## 14 Tennessee               4.39  18.6       32.5          12        497      543
## 15 Alabama                 4.40  17.2       31.1           8        491      538
## 16 Mississippi             4.08  17.5       26.8           4        496      540
## 17 Utah                    3.66  24.3       29.1           4        513      563
## # ... with 15 more variables: sat1995 <int>, region <fct>, pop <int>,
## #   verbal1992 <int>, math1992 <int>, percent1992 <int>,
## #   pupilexpense1992 <dbl>, salary1992 <int>, sat1992 <int>, averagesat <dbl>,
## #   averagepupilexpense <dbl>, averagesalary <dbl>, averagepercent <dbl>,
## #   averagesatverbal <dbl>, averagesatmath <dbl>
```

```
combinedstates %>%
  group_by(state)%>%
  arrange(desc(averagesalary))%>%
  select(ratio)
```

```
## # A tibble: 50 x 2
## # Groups:   state [50]
##    state         ratio
##    <fct>         <dbl>
##  1 Connecticut    14.4
##  2 Alaska         17.6
##  3 New York       15.2
##  4 New Jersey     13.8
##  5 Pennsylvania   17.1
##  6 California     24
##  7 Michigan       20.1
##  8 Maryland       17
##  9 Rhode Island   14.7
## 10 Massachusetts  14.8
## # ... with 40 more rows
```

```
combinedstates %>%
  group_by(state)%>%
  arrange(desc(ratio))%>%
  select(pop)
```

```
## # A tibble: 50 x 2
## # Groups:   state [50]
##    state        pop
##    <fct>      <int>
##  1 Utah        1723
##  2 California 29760
##  3 Washington  4867
##  4 Michigan    9295
##  5 Oregon      2842
##  6 Arizona     3665
##  7 Florida    12938
##  8 Idaho       1007
##  9 Nevada      1202
## 10 Tennessee   4877
## # ... with 40 more rows
```

```
#arranged the data by various variables to see what states fared best (grouped by states), also selecting and filtering vari
ous variables to explore dataset
combinedstates_num <- combinedstates %>%
  select_if(is.numeric)
cor(combinedstates_num, use = "pairwise.complete.obs")%>%
  kbl%>%
  kable_styling()
```

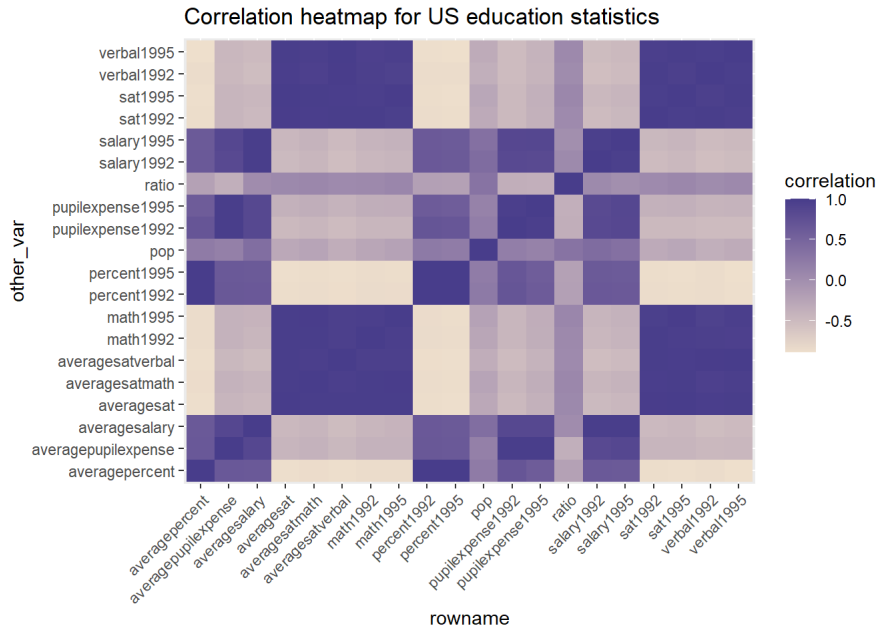
	pupilexpense1995	ratio	salary1995	percent1995	verbal1995	math1995	sat1995	pop	verbal1992
pupilexpense1995	1.0000000	-0.3710254	0.8698015	0.5926274	-0.4100499	-0.3494141	-0.3805370	0.1434879	-0.4209078
ratio	-0.3710254	1.0000000	-0.0011461	-0.2130536	0.0637666	0.0954217	0.0812538	0.3192271	0.0290800
salary1995	0.8698015	-0.0011461	1.0000000	0.6167799	-0.4769636	-0.4013128	-0.4398834	0.3531944	-0.5003208
percent1995	0.5926274	-0.2130536	0.6167799	1.0000000	-0.8932630	-0.8693839	-0.8871187	0.2137348	-0.8611290
verbal1995	-0.4100499	0.0637666	-0.4769636	-0.8932630	1.0000000	0.9702560	0.9915033	-0.3282947	0.9781740
math1995	-0.3494141	0.0954217	-0.4013128	-0.8693839	0.9702560	1.0000000	0.9935024	-0.2249310	0.9386033
sat1995	-0.3805370	0.0812538	-0.4398834	-0.8871187	0.9915033	0.9935024	1.0000000	-0.2752100	0.9642335
pop	0.1434879	0.3192271	0.3531944	0.2137348	-0.3282947	-0.2249310	-0.2752100	1.0000000	-0.3668112
verbal1992	-0.4209078	0.0290800	-0.5003208	-0.8611290	0.9781740	0.9386033	0.9642335	-0.3668112	1.0000000
math1992	-0.3581530	0.0633206	-0.4232865	-0.8631369	0.9550743	0.9740209	0.9724050	-0.2650858	0.9617535
percent1992	0.6103693	-0.2059106	0.6365475	0.9968855	-0.8844948	-0.8524616	-0.8739032	0.2393264	-0.8577610
pupilexpense1992	0.9684576	-0.3618375	0.8628594	0.6788224	-0.5022311	-0.4478127	-0.4767492	0.1897784	-0.5032199
salary1992	0.8242468	0.0542548	0.9656808	0.6297318	-0.5096810	-0.4310896	-0.4712656	0.4041393	-0.5380999
sat1992	-0.3916596	0.0475418	-0.4642589	-0.8705263	0.9753724	0.9664893	0.9779009	-0.3163477	0.9893672
averagesat	-0.3877982	0.0660811	-0.4536147	-0.8843412	0.9895176	0.9864848	0.9952956	-0.2957974	0.9812193
averagepupilexpense	0.9923383	-0.3694298	0.8732987	0.6400936	-0.4590276	-0.4009909	-0.4312780	0.1675850	-0.4650787
averagesalary	0.8558095	0.0250172	0.9924431	0.6282380	-0.4965493	-0.4188543	-0.4585172	0.3803199	-0.5224999
averagepercent	0.6014475	-0.2098535	0.6265732	0.9993094	-0.8898255	-0.8620858	-0.8815810	0.2259592	-0.8602111
averagesatverbal	-0.4173753	0.0478858	-0.4905038	-0.8831180	0.9952625	0.9607577	0.9841721	-0.3481141	0.9937419
averagesatmath	-0.3557169	0.0812486	-0.4140531	-0.8721661	0.9695773	0.9945387	0.9902487	-0.2449019	0.9553845

```
#created a correlation matrix of all numeric variables
```

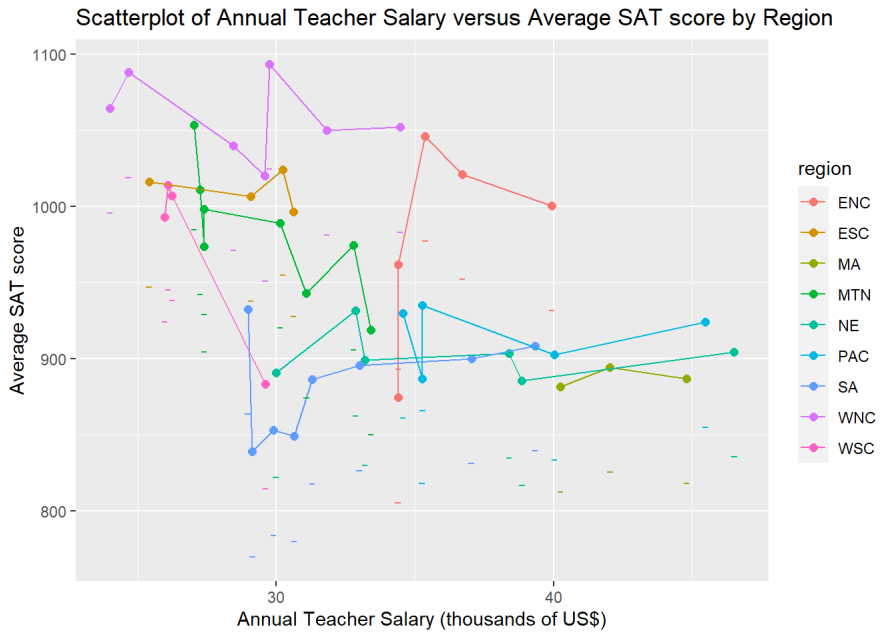
The statistics found in this section are quite interesting. There are so many, however, that it would be hard to discuss them all without writing a whole page that would seem quite redundant. One of the main things I noticed across the summaries was that there is quite a bit of disparity across the states. Among individual states, the average SAT score had a range of more than 200 points, the average salary of teachers had a range of more than $20,000, and the average percent of graduating students varied by more than 70 percentange points, to name a few. Another trend that could be seen across the data was that Northern states and regions had better educational statistics, such as lower pupil to teacher ratios, higher expenditures per pupil, and typically higher SAT scores.

VISUALIZATION

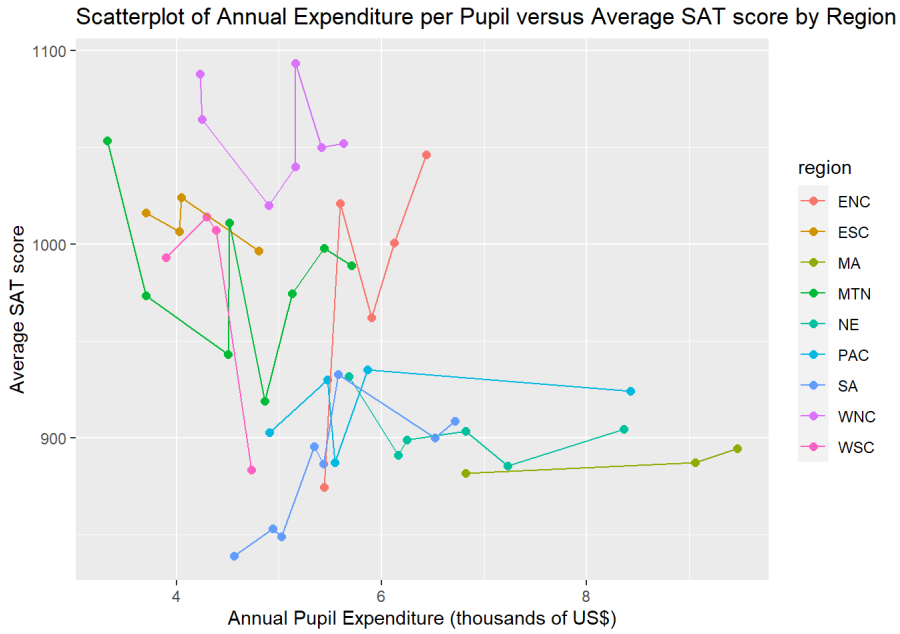
```
library(ggplot2)
combinedstates_num2 <- cor(combinedstates_num, use = "pairwise.complete.obs") %>%
  as.data.frame %>%
  rownames_to_column %>%
  pivot_longer(-1, names_to = "other_var", values_to = "correlation")
ggplot(combinedstates_num2, aes(rowname, other_var, fill= correlation)) +
  geom_tile()+
  scale_fill_gradient(low = "antiquewhite2", high = "darkslateblue")+
  labs(title = "Correlation heatmap for US education statistics")+
  theme(axis.text.x = element_text(angle=45, hjust=1))
```




```
#heatmap created with angled x axis, looks at all variables and their correlation
combinedstates<- combinedstates%>%
  mutate(sd_averagesat = sd(averagesat))
ggplot(data = combinedstates, aes(x = averagesalary, y = averagesat, color = region,stat="summary", fun="mean")) +
  geom_point(size = 2) +
  geom_line()+
  labs(title = "Scatterplot of Annual Teacher Salary versus Average SAT score by Region", y = "Average SAT score", x = "Annu
al Teacher Salary (thousands of US$)")+
  geom_errorbar(aes(ymin=averagesat-sd_averagesat, ymax=averagesat-sd_averagesat), width=.2,)
```




```
#scatterplot of salary versus sat score grouped by region, error bars were attempted but serve only to confuse.
ggplot(data = combinedstates, aes(x = averagepupilexpense, y = averagesat, color = region)) +
  geom_point(size = 2) +
  geom_line()+
  labs(title = "Scatterplot of Annual Expenditure per Pupil versus Average SAT score by Region", y = "Average SAT score", x
 = "Annual Pupil Expenditure (thousands of US$)")
```




```
#scatterplot used again as I only have one categorical variable for the whole dataset, and the only really dependent variabl
e that shows the education level of each state is the average SAT score, so this will go on the y label once more. no error
 bars as they are cluttering.
```

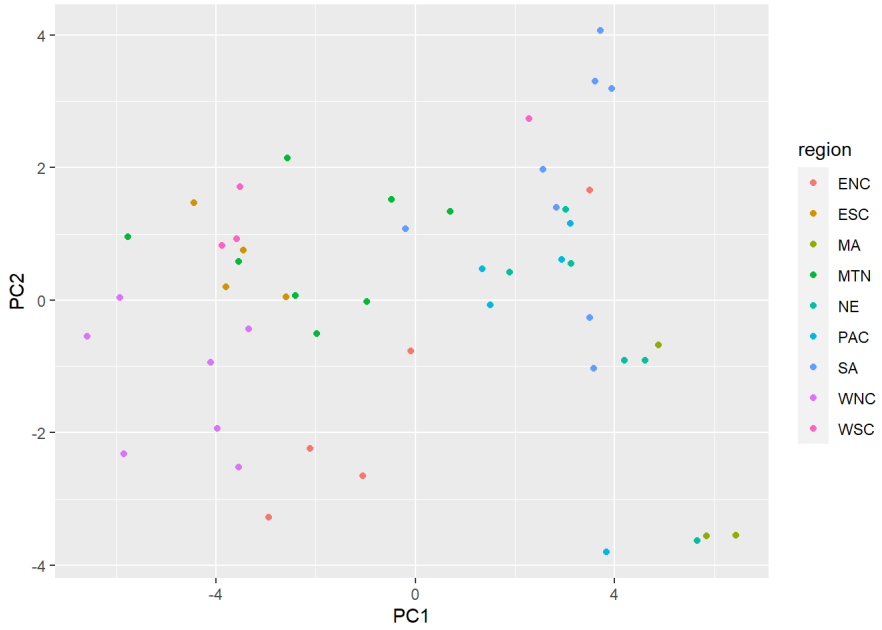
The correlation heatmap shows a multitude of variables, and there are strong correlations between verbal and math SAT scores, and in turn the average SAT score. However, the biggest take away from this plot is that there are not strong correlations between any of the variables. This trend continues with the following two scatterplots, which relate teacher salaries and pupil expenditures to SAT score by region. Neither plot shows a strong correlation, and if anything an increase in taecher salary seems to either have no substantial effect on SAT scores, or may even have a negative correlation. Again, the data shows no strong correlation with any variables and SAT scores, which are a sign of good education.

PCA/CLUSTERING

```
pca <- combinedstates_num%>%
  scale()%>%
  prcomp()
#standardized and scaled already solely numeric dataset
names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

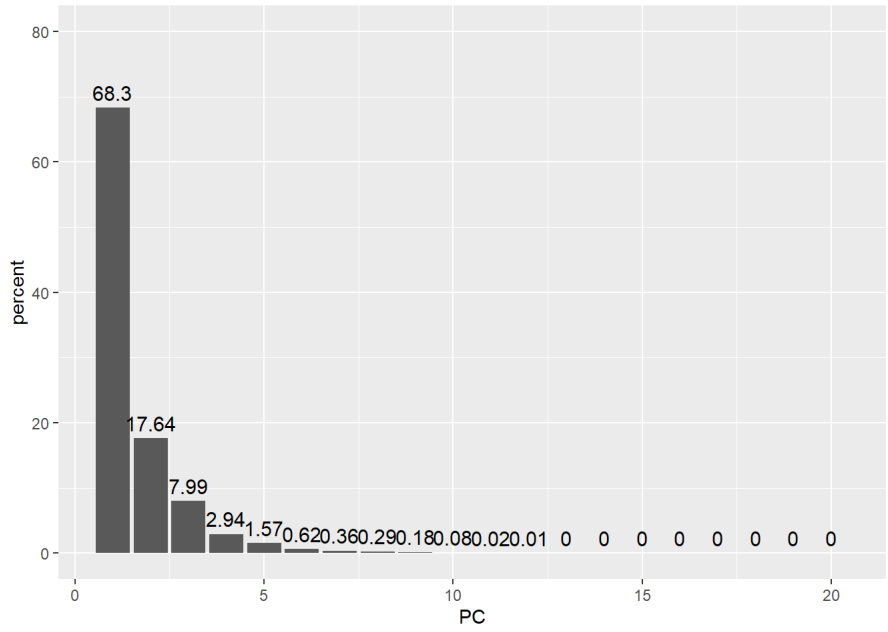
```
pca_data <- data.frame(pca$x, region = combinedstates$region, state = combinedstates$state)
#added back in categorical variables
ggplot(pca_data, aes(x = PC1, y = PC2, color = region)) +
  geom_point()
```




```
#mapped on scatterplot the principal components by region
percent <- 100* (pca$sdev^2 / sum(pca$sdev^2))
percent
```

```
##  [1] 6.829840e+01 1.763876e+01 7.991000e+00 2.940537e+00 1.570842e+00
##  [6] 6.192295e-01 3.624902e-01 2.878064e-01 1.839556e-01 8.286377e-02
## [11] 1.535115e-02 8.763822e-03 1.060091e-30 4.664511e-31 1.752995e-31
## [16] 1.142076e-31 6.346171e-32 4.844422e-32 3.187157e-32 2.770323e-32
```

```
#saw how much effect components had on variance within data
perc_data <- data.frame(percent = percent, PC = 1:length(percent))
ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col() +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -0.5) +
  ylim(0, 80)
```




```
#visualize with a bargraph the variance caused by PCs
```

The first principal component, which holds the most information out of all the principal components, accounts for almost 70% of the variance in the dataset (68.3%), and the second accounts for about 18% (17.64%). Because of the many redundant variables in my dataset, such as the different years for variables and the different types of SAT scores, this analysis was useful in better understanding and simplifying the data. I cannot identify any specific patterns in these PCA clustering graphs.

```
##           sysname           release           version          nodename
##         "Windows"          "10 x64"     "build 19042" "LAPTOP-VDO7L3JC"
##           machine             login              user    effective_user
##          "x86-64"             "HP"              "HP"              "HP"
```